Approximation de lois impropres et applications

Christèle Bioche

N° d'Ordre : D.U.2626

UNIVERSITÉ BLAISE PASCAL
U.F.R. Sciences et Technologies

ÉCOLE DOCTORALE DES SCIENCES FONDAMENTALES

THÈSE

présentée pour obtenir le grade de

DOCTEUR D'UNIVERSITÉ

Spécialité :
MATHÉMATIQUES APPLIQUÉES

Par **Christèle BIOCHE**

Approximation de lois impropres et applications

Soutenue publiquement le 27 novembre 2015

après avis de :

Brunero LISEO	Professeur	Université de Rome 1
Judith ROUSSEAU	Professeur	Université Paris Dauphine

devant la commission d'examen composée de :

Pierre DRUILHET	Professeur	Université Blaise Pascal	(Directeur)
Brunero LISEO	Professeur	Université de Rome 1	(Rapporteur)
Jean-Michel MARIN	Professeur	Université de Montpellier	
Anne PHILIPPE	Professeur	Université de Nantes	
Laurent SERLET	Professeur	Université Blaise Pascal	
Anne-Françoise YAO	Professeur	Université Blaise Pascal	

« Si loin que vous alliez, si haut que vous montiez, il vous faut commencer par un simple pas. »

Shitao

Remerciements

Je tiens tout d'abord à remercier Pierre Druilhet sans qui ce travail n'aurait pas eu lieu. Merci d'avoir répondu positivement à ma demande de stage de M2, de m'avoir fait découvrir les statistiques bayésiennes et de m'avoir fortement encouragée à entamer cette thèse. Merci aussi pour ta disponibilité, ta gentillesse et ta confiance en moi, même lorsque j'en manquais moi-même.

Je souhaite aussi remercier Judith Rousseau et Brunero Liseo (Grazie Brunero) de m'avoir fait l'honneur de lire mon travail et d'avoir accepté d'être les rapporteurs de cette thèse. De plus, je remercie Anne Philippe, Anne-Françoise Yao, Jean-Michel Marin et Laurent Serlet d'avoir accepté d'être présents au sein de mon jury.

Merci à tous les membres du laboratoire de Clermont-Ferrand que j'ai côtoyés pendant ces trois années. En particulier, je remercie Catherine Savona, Thierry Buffard, Erwann Saint Loubert Bié et Thierry Lambre avec lesquels j'ai été amenée à travailler pour mes enseignements ou des projets de vulgarisation des mathématiques. Je tiens aussi à remercier Stéphanie Léger pour nos discussions aussi bien professionnelles que personnelles, tes conseils m'ont été précieux. Enfin Annick, Karine, Laurence, Marie-Paule et Valérie : les petites mamans du labo ; merci pour votre aide au sein du laboratoire mais aussi pour votre gentillesse et votre bienveillance.

J'adresse aussi mes remerciements aux membres du laboratoire de Nantes pour votre accueil chaleureux et l'aide que certains m'ont apportée dans la préparation de ma soutenance.

Je tiens à remercier les doctorants qui m'ont accompagnée pendant ces trois années. Un merci plus particulier à Colin et Romuald pour tous ces déjeuners partagés, ces pauses café rallongées et tous vos conseils. Vous m'avez bien manqué pendant cette dernière année ! J'en profite pour remercier aussi Muriel, pour sa gentillesse et sa bonne humeur permanente. Je remercie aussi Lorena, Thérèse et Honoré, je garde un très bon souvenir de notre semaine de formation à Bellenaves.

Un très grand merci à Audrey, mon acolyte de thèse. Merci pour ton soutien en toute situation, pour tous ces fous rires, ces joggings, ces gâteaux, ces petits plats

du midi, ces soupes du mardi soir, ces heures à discuter... Et surtout un grand merci pour ton hospitalité pendant mon dernier été de thèse.

Merci à tous ceux qui ont rendu mes années à Clermont si agréables. Notamment Susana et Adrien pour les nombreuses sorties les week-ends ; Oonalee pour toutes ces soirées à discuter de Vital ou de Vital Food suivant le dernier sorti ; les filles du service de santé publique pour tous ces déjeuners remplis de bonne humeur et de nombreuses (très nombreuses) tablettes de chocolat et Élise et Alex pour la qualité de vos cours et l'équilibre qu'ils m'ont apporté.

Laure, Ingrid, Caroline, Marie-Aude, Mathilde, Tiphaine, Caroline, Charlotte, Céline, Gaspard, Damien et Marjo ; vous êtes nombreux à avoir fait honneur à la chambre d'amis de mon appartement clermontois pendant ces trois années et ça m'a fait très plaisir ! Je tiens à vous remercier pour vos encouragements, la confiance que certains ont manifestée en mon travail et les bons moments passés avec vous tous.

Merci à tout le clan Bioche-Raberin, du petit neveu aux grands-parents en passant par les parents, beaux-parents, frères, belles-soeurs ; merci pour tout le soutien que vous m'avez apporté, ainsi que pour vos efforts pour comprendre le monde de la recherche. Que ce soit en essayant de comprendre ce que je faisais, en apprenant juste le titre de ma thèse ou en m'en parlant le moins possible ; vous avez su m'entourer de la meilleure des façons durant ces trois années.

Je tiens à remercier plus particulièrement mes parents de m'avoir donné le goût du travail.

Je remercie enfin Jean-Louis d'être à mes côtés depuis ces neuf longues années d'étude. Merci pour ta patience, qui n'est pourtant pas ta qualité première, et pour ton soutien.

Table des matières

Plan de la thèse

Le premier chapitre de cette thèse est une présentation générale des statistiques bayésiennes. Nous évoquons quelques domaines d'application, présentons le paradigme bayésien puis discutons du point crucial de l'analyse bayésienne : le choix de la distribution *a priori*. Notamment, nous exposons les distributions non-informatives les plus classiques. Ces distributions étant le plus souvent impropres, nous achevons ce premier chapitre en discutant sur l'utilisation de telles distributions. Certains auteurs déconseillent l'utilisation de telles lois *a priori* mais celles-ci présentent tout de même de nombreux avantages.

Dans le deuxième chapitre, nous présentons les travaux de différents auteurs visant à justifier l'utilisation des distributions *a priori* impropres. Une première approche consiste à revisiter les fondements des probabilités. Une autre, sur laquelle nous nous attarderons davantage, consiste à faire apparaitre les distributions *a priori* impropres comme limites naturelles de lois *a priori* propres. Nous verrons que la vraisemblance du modèle intervient dans tous les modes de convergence proposés dans la littérature. Une question se pose alors : la limite d'une suite de distributions *a priori* dépend-elle du modèle statistique ?

Le troisième chapitre contient notre premier article *Approximation of improper prior* à paraître dans *Bernoulli Journal*. Le but de cet article est de définir un mode de convergence sur les suites d'*a priori* qui soit intrinsèque ; c'est-à-dire indépendant du modèle statistiques. La quasi-totalité des *a priori* usuels étant des mesures de Radon strictement positives, nous définissons un mode de convergence sur cet ensemble. Ce mode de convergence, que nous appelons *convergence q-vague*, est indépendant du modèle statistique. Nous démontrons que pour ce mode de convergence, tout *a priori* impropre peut être approximé par une suite d'*a priori* propres et inversement. Nous étudions quelques propriétés de ce mode de

convergence et les convergences induites sur les distributions ou estimateurs *a posteriori* lorsque l'on suppose la convergence q-vague des *a priori*. Enfin, ce mode de convergence permet de comprendre l'origine du paradoxe de Jeffreys-Lindley. Nous proposons ensuite une partie contenant quelques résultats complémentaires qui n'apparaissent pas dans l'article.

Dans le quatrième chapitre, nous utilisons les résultats obtenus grâce à la convergence q-vague pour fournir des recommandations sur le choix des *a priori* dans le cadre du removal sampling. Nous commençons par exposer la méthode de removal sampling, sa modélisation et les techniques usuelles d'estimation utilisées dans ce cadre. Puis, nous proposons un article, *Bayesian estimation of abundance by removal sampling*, dans lequel nous étudions de manière théorique les propriétés du modèle associé au removal sampling. Nous établissons des conditions nécessaires et suffisantes sur les *a priori* pour obtenir des estimateurs *a posteriori* bien définis. Enfin, nous montrons à l'aide de la convergence q-vague, que l'utilisation d'*a priori* vagues n'est pas adaptée car les estimateurs obtenus montrent une grande dépendance aux hyperparamètres.

Le cinquième chapitre est une ébauche d'article. Nous cherchons des conditions sur les *a priori* pour obtenir la convergence logarithmique des *a posteriori*. Nous introduisons un nouveau mode de convergence sur les *a priori*, un peu plus restrictif que la convergence q-vague que nous appelons convergence q-monotone. La convergence q-monotone des *a priori* implique la convergence logarithmique des *a posteriori*. Ceci généralise le résultat de Berger et al. (2009) qui n'avaient travaillé que sur des suites d'*a priori* obtenues par troncature. Nous généralisons aussi le résultat qu'ils proposent sur la convergence en espérance logarithmique des *a posteriori* dans le cadre du modèle de position en l'étendant à d'autres types de suites approximantes que les suites obtenues par troncature.

Chapitre 1

Les Statistiques bayésiennes

Cette première partie, fortement inspirée de *The bayesian choice* (Robert, 2007), présente les statistiques bayésiennes. Nous donnons d'abord un aperçu général de quelques domaines d'application et de la méthode bayésienne. Puis, nous nous intéressons au point crucial de l'approche bayésienne : le choix de la distribution *a priori*. Nous présentons ensuite quelques méthodes de construction d'*a priori* non-informatifs. Enfin, nous nous concentrons sur les distributions *a priori* impropres et les difficultés liées à leur utilisation.

1.1 Introduction

La méthode bayésienne est un ensemble de techniques statistiques utilisées pour modéliser des problèmes, extraire de l'information de données brutes et prendre des décisions de façon cohérente et rationnelle. Son cadre d'application est général, mais ses avantages sont déterminants lorsque l'information disponible est incertaine ou incomplète. Bien que les premiers travaux d'inspiration bayésienne datent du XVII$^{\text{ème}}$ siècle, cette méthode connait un regain de popularité depuis quelques décennies. Ce renouveau est sensible dans des domaines très variés, en partie grâce à la disponibilité de calculateurs puissants, mais aussi à une évolution de la pensée statistique et des problèmes abordés.

Les statistiques bayésiennes sont très utilisées en sciences sociales et politiques, car les données y sont rares et coûteuses à collecter (Gelman et al., 2004). Elles

servent aussi en physique des particules (Cousins, 1995; Demortier, 2006), en thermodynamique (Chatterjee et al., 1998), en mécanique statistique (Jaynes, 1957), en chimie (Vines et al., 1993; Pohorille and Darve, 2006), en génétique (Smyth, 2004; Chan et al., 2006), et en bioinformatique (Wilkinson, 2007).

La méthode bayésienne est également employée en sciences cognitives, pour modéliser les comportements animaux et humains comme des prises de décisions rationnelles (Kording, 2004). Les neurosciences computationnelles ont pour but de comprendre le fonctionnement des neurones en tant que systèmes de traitement de l'information optimaux. L'approche bayésienne y est aussi prometteuse (Pouget et al., 2003; Wu et al., 2003; Deneve, 2005).

En intelligence artificielle, la proposition de Bessière et al. (1998a,b) d'une théorie probabiliste des systèmes cognitifs sensi-moteurs a conduit à une méthode de programmation bayésienne des robots (Lebeltel et al., 2003).

Tous ces travaux reposent sur la contribution fondamentale de Jaynes résumée dans son livre posthume *Probability Theory : The Logic of Science* (Jaynes, 2003).

1.2 Le modèle statistique

Nous ne considérons dans cette thèse que l'approche paramétrique. Nous supposons donc que les observations $x_1, \ldots, x_n$, sur lesquelles l'analyse statistique se fonde, proviennent de lois de probabilité paramétriques. Ainsi, $x_i (1 \leqslant i \leqslant n)$ a une distribution de densité $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1})$ sur $\mathbb{R}^p$, telle que le paramètre θ_i soit inconnu et la fonction f_i soit connue. Ce modèle peut être représenté par $x \sim f(x|\theta)$ où x est le vecteur d'observations et θ l'ensemble des paramètres $\theta_1, \ldots, \theta_n$, éventuellement tous égaux. Le vecteur θ est toujours de dimension finie. Cette représentation est unificatrice dans le sens où elle aborde de manière similaire une observation isolée, des observations dépendantes, et des observations indépendantes et identiquement distribuées (iid) $x_1, \ldots, x_n$ de même loi, $f(x_1|\theta)$. Dans le dernier cas, $x = (x_1, \ldots, x_n)$ et

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Une fois le modèle statistique identifié, l'objectif principal de l'analyse statistique est de nous conduire à une inférence sur le paramètre θ. Nous utilisons l'observation de x pour améliorer notre connaissance du paramètre θ.

1.3 Le paradigme bayésien

1.3.1 La Formule de Bayes

Soient A et B deux événements aléatoires tels que $P(B) \neq 0$. La probabilité de A conditionnellement à la réalisation de B est, par définition, donnée par la relation suivante :

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

où $P(A,B)$ est la probabilité que les deux événements A et B aient lieu simultanément.

Puisque $P(A,B) = P(B,A)$, alors les deux probabilités conditionnelles $P(A|B)$ et $P(B|A)$ sont reliées par :

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}.$$

Cette relation appelée *Théorème de Bayes* est en fait un principe d'actualisation : elle décrit la mise à jour de la vraisemblance de A de $P(A)$ vers $P(A|B)$ une fois que B a été observé. Bayes (1763) donne une version continue de ce résultat : pour deux variables aléatoires x et y, de distributions conditionnelle $f(x|y)$ et marginale $g(y)$, la distribution conditionnelle de y sachant x est

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\mathrm{d}y}.$$

Ce théorème d'inversion est naturel d'un point de vue probabiliste mais Bayes et Laplace sont allés plus loin et ont considéré que l'incertitude sur le paramètre θ d'un modèle peut être décrite par une distribution de probabilité π sur Θ appelée distribution *a priori*. L'inférence est alors fondée sur la distribution de θ

conditionnelle à x, $\pi(\theta|x)$, appelée distribition *a posteriori* et définie par

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}. \tag{1.1}$$

Cette équation sera appelée la *Formule de Bayes*. Le dénominateur étant indépendant de θ, la Formule de Bayes peut s'écrire de la façon suivante :

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta). \tag{1.2}$$

Le passage de la distribution *a priori* à la distribution *a posteriori* des paramètres du modèle peut être interprété comme une mise à jour de la connaissance sur la base des observations.

1.3.2 La méthode d'analyse bayésienne

La méthode bayésienne pour résoudre un problème d'analyse de données est décrite informellement par les étapes suivantes :

1. Formuler notre connaissance *a priori* du problème.
   (a) Établir un modèle $\mathcal{M}$ probabiliste de la génération des données.
   (b) Si ce modèle possède des paramètres libres θ, traduire nos connaissances en une distribution de probabilité sur ces paramètres.
2. Collecter les données.
3. Utiliser la Formule de Bayes (1.1) pour mettre à jour nos connaissances.
4. Utiliser la distribution *a posteriori* obtenue pour :
   (a) faire des prédictions,
   (b) construire des intervalles de confiance,
   (c) construire des tests,
   (d) commencer une nouvelle analyse pour laquelle cette connaissance *a posteriori* deviendra notre nouvelle connaissance *a priori*,
   (e) etc.
5. Vérifier la pertinence des résultats. Si les conclusions sont manifestement fausses, retourner au point 1.

En toute rigueur, le modèle et les distributions *a priori* doivent être établis indépendamment des données et l'*a priori* doit vraiment représenter un état des connaissances réelles du statisticien. En pratique, il ne sera pas toujours possible de respecter ces contraintes. Il faudra alors prendre garde à ce que les libertés d'approximation décidées n'influent pas trop sur les résultats.

1.4 Le choix de la distribution *a priori*

Choisir la loi *a priori* revient à traduire le savoir de l'expert sur le paramètre en une distribution de probabilité. Le choix de la loi *a priori* est une étape fondamentale dans l'analyse bayésienne. En effet, une fois que cette loi est connue l'inférence peut être menée de manière quasi-systématique.

Dans la pratique, il est rare que l'information *a priori* soit suffisamment précise pour conduire à une détermination exacte d'une loi *a priori*. Le statisticien est donc amené à faire un choix arbitraire de loi *a priori*, ce qui peut modifier considérablement l'inférence qui en découle. Ce choix peut avoir différentes motivations, les stratégies sont diverses. Elles peuvent se baser sur des expériences du passé ou sur une intuition, une idée que le praticien a du phénomène aléatoire qu'il est en train de suivre. Elles peuvent être également motivées par des aspects de calculabilité. Enfin, ces stratégies peuvent aussi tenir compte du fait que l'on ne sait rien par l'utilisation de lois non-informatives. Certaines situations requièrent une détermination partiellement automatisée de la loi *a priori* comme dans le cas extrême où l'information *a priori* est complètement absente.

Cette étape, qui est la clé de voûte de l'analyse bayésienne est aussi celle à laquelle l'approche bayésienne doit toutes ses critiques. En effet, les détracteurs de l'approche bayésienne attirent l'attention sur le fait qu'il n'y a pas une façon unique de choisir une loi *a priori*, et que ce choix a un impact sur l'inférence résultante.

1.5 Distributions *a priori* non informatives

Certains auteurs ont tenté d'introduire des *a priori* ne dépendant pas de l'état de connaissance d'un agent mais déduits de règles formelles. Ces lois *a priori* non-informatives représentent une ignorance sur le problème considéré, mais ne signifient pas que l'on ne sache absolument rien sur la distribution statistique du paramètre. Ce sont des lois qui portent une information sur le paramètre à estimer dont le poids dans l'inférence est réduit.

Ces *a priori* ont des avantages : ils sont faciles à formuler, ils donnent l'apparence de l'objectivité, ils nous évitent de travailler avec des *a priori* subjectifs mal formulés, ils possèdent des propriétés analytiques agréables et ils ont de bonnes propriétés fréquentistes. Les différentes méthodes proposées pour obtenir ce type d'*a priori* ont pour point commun de n'utiliser comme source d'information que la forme de la fonction de vraisemblance $f(x|\theta)$ définie par le modèle.

Les lois non-informatives peuvent être considérées comme des lois de référence auxquelles chacun pourrait avoir recours quand toute information *a priori* est absente ou minime. Certaines de ces lois sont plus utiles ou plus efficaces que d'autres mais ne peuvent être perçues comme moins informatives que d'autres. Il est désormais largement admis qu'il n'existe pas d'*a priori* absolument non-informatif (Kass and Wasserman, 1996).

Nous décrivons maintenant quelques unes des techniques les plus importantes de construction de lois non-informatives.

1.5.1 Distribution *a priori* de Laplace

Laplace fût le premier à utiliser des techniques non-informatives puisque, bien que ne disposant pas d'information *a priori* pour les paramètres qu'il étudiait, il munit ces paramètres d'une loi qui prend en compte son ignorance en donnant la même vraisemblance à chaque valeur possible, soit donc en utilisant une loi uniforme. Son raisonnement, appelé plus tard *principe de la raison insuffisante*, se fondait sur l'équiprobabilité des événements élémentaires.

Trois critiques ont été plus tard avancées sur ce choix. Premièrement, les lois résultantes sont impropres quand l'espace des paramètres n'est pas compact et

certains statisticiens se refusent à utiliser de telles lois car elles mènent à des difficultés comme nous le verrons dans la Section 1.6. Deuxièmement, le principe des événements équiprobables n'est pas cohérent en terme de partitionnement. Si $\Theta = \{\theta_1, \theta_2\}$, la règle de Laplace donne $\pi(\theta_1) = \pi(\theta_2) = \frac{1}{2}$ mais si la définition de Θ est plus détaillée avec $\Theta = \{\theta_1, \omega_1, \omega_2\}$, la règle mène à $\pi(\theta_1) = \frac{1}{3}$, ce qui est incohérent avec la première formulation. On peut passer outre ce problème de cohérence en déclarant que le niveau de partitionnement doit être fixé à un certain stade de l'analyse et que l'introduction d'un degré plus fin dans le partitionnement modifie le problème d'inférence. La troisième critique est plus fondamentale, elle concerne le problème d'invariance par reparamétrisation. Si on passe de $\theta \in \Theta$ à $\eta = g(\theta)$ par une transformation bijective g, l'information *a priori* reste totalement inexistante et ne devrait pas être modifiée. Cependant, si $\pi(\theta) = 1$, la loi *a priori* sur η est :

$$\pi^*(\eta) = \left| \frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta) \right|$$

par la formule de changement de variable. Donc $\pi^*(\eta)$ est le plus souvent non constante.

1.5.2 Distributions *a priori* invariantes

L'idée est de rechercher des *a priori* invariants sous l'action d'un certain groupe de transformations afin d'obtenir une loi non-informative compatible avec les exigeances d'invariance. Cette méthode nous pousse à considérer la mesure de Haar à droite du groupe agissant sur l'ensemble des paramètres (Eaton, 1989; Kass and Wasserman, 1996).

L'approche invariante n'est que partiellement satisfaisante car elle implique la référence à une structure d'invariance qui peut parfois être choisie de plusieurs manières, ne pas exister, ou être sans intérêt pour le décideur.

1.5.3 Distributions *a priori* de Jeffreys

Les lois non-informatives de Jeffreys (1946, 1961) sont fondées sur l'information de Fisher donnée par

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \right]$$

dans le cas unidimensionnel. Sous certaines conditions de régularité, cette information est aussi égale à

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right]. \tag{1.3}$$

La loi *a priori* de Jeffreys est

$$\pi(\theta) \propto I^{\frac{1}{2}}(\theta)$$

définie à un coefficient de renormalisation près quand π est propre. La loi *a priori* de Jeffreys est invariante par reparamétrisation puisque pour une transformation bijective donnée h qui transforme le paramètre θ en $h(\theta)$, nous avons la transformation jacobienne

$$I(\theta) = I(h(\theta))(h'(\theta))^2.$$

Le choix d'une loi *a priori* dépendant de l'information de Fisher se justifie par le fait que $I(\theta)$ est largement accepté comme un indicateur de la quantité d'information apportée par le modèle ou l'observation sur θ (Fisher, 1956). Il paraît intuitivement justifié que les valeurs pour lesquelles l'information de Fisher est plus grande doivent être plus probables *a priori* car ceci équivaut à minimiser l'influence de la loi *a priori* qui est donc aussi non-informative que possible.

Dans le cas où θ est un paramètre multidimensionnel, on définit la matrice d'information de Fisher par généralisation de l'équation (1.3). Pour $\theta \in \mathbb{R}^k$, $I(\theta)$ a les éléments suivants :

$$I_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\theta) \right], (i, j = 1, \ldots, k).$$

La loi non-informative de Jeffreys est alors définie par :

$$\pi(\theta) \propto [\det(I(\theta))]^{\frac{1}{2}}$$

elle est toujours invariante par reparamétrisation.

L'approche de Jeffreys fournit une des meilleures techniques automatiques pour obtenir une loi *a priori* non-informative. De plus, elle permet de retrouver les estimateurs classiques. Cependant, elle a été critiquée par certains Bayésiens comme étant un outil sans justification subjective en terme d'information *a priori*.

1.5.4 Distributions *a priori* de référence

Bernardo propose une modification de l'approche de Jeffreys en présentant les distributions de référence. Une différence majeure est que cette méthode fait la distinction entre paramètre d'intérêt et paramètre de nuisance. Par conséquent, la loi résultante ne dépend pas seulement de la loi d'échantillonnage, mais aussi du problème inférentiel considéré.

Quand $x \sim f(x|\theta)$ et $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt, la loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)\mathrm{d}\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x|\theta_1)$.

Cette stratégie peut se généraliser si $\theta = (\theta_1, \ldots, \theta_n)$, et si on ordonne les θ_i par intérêt croissant.

La méthode se justifie comme fournissant la loi *a priori* qui maximise l'information *a posteriori* (Bernardo, 1979a; Berger and Bernardo, 1992).

1.6 Distributions *a priori* impropres

Lorsque le paramètre θ peut être traité comme une variable aléatoire avec une distribution de probabilité Π connue, nous avons vu que le théorème de Bayes

est la base de l'inférence bayésienne car il donne la distribution *a posteriori*. Cependant, dans de nombreux cas, la distribution *a priori* est déterminée par des critères subjectifs ou théoriques qui conduisent à une mesure infinie sur l'espace des paramètres Θ plutôt qu'à une mesure de probabilité, c'est-à-dire à une mesure Π telle que

$$\int_{\Theta} \pi(\theta) \mathrm{d}\theta = +\infty.$$

Dans de tels cas, on dit que la distribution *a priori* est impropre. Ce type de loi n'a donc plus d'intérêt que calculatoire et s'interprète difficilement comme le fait remarquer Lindley (1990) « L'erreur est de les interpréter [les lois *a priori* non-informatives] comme des représentations d'une complète ignorance ». Quand une telle loi a été obtenue par des méthodes automatiques telles que celles décrites dans la section 1.5, elle paraît plus suceptible aux critiques mais soulignons les points suivants :

1. Ces approches automatiques sont souvent la seule façon d'obtenir une distribution *a priori* dans un cadre non-informatif. Cette généralisation du paradigme bayésien rend ainsi possible une extension supplémentaire de l'applicabilité des techniques bayésiennes.
2. Les performances des estimateurs obtenus à partir de ces distributions généralisées sont en général suffisamment bonnes pour justifier leur utilisation.
3. Une perspective « récente » (Berger, 2000) est que les lois *a priori* impropres devraient être privilégiées par rapport aux lois *a priori* propres vagues, comme une distribution $\mathcal{N}(0, 100^2)$, car ces dernières donnent une fausse impression de sécurité due à leur caractère propre tout en manquant de robustesse en terme d'influence sur les résultats d'inférence.
4. Les lois *a priori* généralisées se situent souvent à la limite des distributions propres.

Nous reviendrons plus en détails sur cette dernière assertion dans la suite de cette thèse. En effet, le chapitre 3 a pour but de définir un mode de convergence pour lequel les distributions *a priori* impropres apparaissent comme des limites naturelles de distributions *a priori* propres.

D'un point de vue pratique, tant que la distribution *a posteriori* est définie, les méthodes bayésiennes restent applicables. En fait, la notion de mesure condition-

nelle n'est pas clairement définie en théorie de la mesure bien que Hartigan (1983) l'ait préconisée comme une extension. Cependant, la convention est de considérer la distribution *a posteriori* $\pi(\theta|x)$ définie formellement par la Formule de Bayes

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}$$

pourvu que la pseudo-distribution marginale $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$ soit correctement définie. La généralisation a des distributions *a priori* impropres ne devrait donc pas poser de problème au sens où une distribution *a posteriori* correspondant à une loi *a priori* impropre peut être utilisée de la même façon qu'une distribution *a posteriori* normale, quand elles sont bien définies. En réalité, traiter des lois *a priori* impropres comme des lois *a priori* standards peut mener à des procédures inadmissibles (voir par exemple Blackwell (1951) et Stein (1956)). Comme l'affirme Bernardo (1997) « One should not interpret any non-subjective prior as a probability distribution ». Voici deux exemples de dysfonctionnements liés à l'utilisation d'*a priori* impropres.

Paradoxe de marginalisation

Le paradoxe de marginalisation, décrit en premier par Stone and Dawid (1972), est étudié plus précisément par Dawid et al. (1973) qui le présentent à l'aide de l'exemple suivant.

Exemple 1.6.1 (Paradoxe de marginalisation)**.** *Considérons $X_1, \ldots, X_n$ des variables aléatoires indépendantes telles que $X_i \sim \exp(\eta)$ et $X_j \sim \exp(c\eta)$ pour $0 \leqslant i \leqslant \xi < j \leqslant n$ avec c connu. Le paramètre est $\theta = (\eta, \xi)$ où $\xi \in \{1, 2, \ldots, n-1\}$ et nous supposons que la loi a priori vérifie $\pi(\eta, \xi) = \pi(\xi)$. On peut alors montrer que $\pi(\xi|X) = \pi(\xi|Z)$ où $Z = \left(\frac{X_2}{X_1}, \ldots, \frac{X_n}{X_1}\right)$ et $f(Z|\xi, \eta) = f(Z|\xi)$ et pourtant il n'existe pas de $\pi(\xi)$ telle que $\pi(\xi|Z) \propto f(Z|\xi)\pi(\xi)$.*

Dawid et al. (1973), Stone (1976) et Jaynes (1980) proposent des solutions partielles à ce paradoxe. Une explication fondamentale est que la loi *a priori* impropre $\pi(\mathrm{d}\eta, \mathrm{d}\xi) = \pi(\xi)\mathrm{d}\xi\mathrm{d}\eta$ ne correspond pas à la loi pseudo marginale $\pi(\mathrm{d}\xi) = \pi(\xi)\mathrm{d}\xi$.

Inconsistance et structure de groupe

Stone (1976) introduit le phénomène de *strong inconsistency* par l'exemple suivant : pour $X \sim \mathcal{N}(\theta, 1)$, si on considère $\pi(\theta) = \exp(4\theta)\mathrm{d}\theta$ comme *a priori* sur θ ; on obtient $P(\theta > x + 1|x) = \Phi(2)$ alors que $P(\theta > x + 1|\theta) = \Phi(-2)$. Cet exemple peut paraître académique mais par la suite Stone (1976) reprend deux exemples, déjà présentés dans un précédent article (Stone, 1970), pour lesquels des *a priori* uniformes mènent au phénomène de *strong inconsistency*. Le premier est en fait une adaptation de l'exemple 11 de Lehmann (1959) (p.24). Les deux exemples mettent en évidence certaines inconsistances provenant de l'utilisation d'*a priori* impropres. Ces inconsistances ne sont cette fois-ci pas liées à un problème de marginalisation mais à un problème de théorie des groupes. En effet, elles ont lieu car les groupes concernés, ici le groupe libre à deux générateurs et le groupe général linéaire 2×2, sont non-moyennables. Un groupe moyennable étant un groupe topologique localement compact que l'on peut munir d'une opération de moyenne sur les fonctions bornées, invariante par les translations par les éléments du groupe (Greenleaf, 1969). Stone (1976) conclut son article en évoquant son septicisme face à l'affirmation de Box and Tiao (1973) qui dit que si l'on utilise des distributions *a priori* impropres pour des cas pratiques on a pas à se soucier de difficultés théoriques.

La solution habituelle, pour éviter ce types de problèmes liés à l'utilisation d'*a priori* impropres, est de déterminer la réponse impropre comme une limite définie à partir d'*a priori* propres.

Chapitre 2

Vers la légitimation des lois *a priori* impropres

D'après les fondements des statistiques bayésiennes, associés à Ramsey, de Finetti et Savage (mais pas Jeffreys), les lois *a priori* impropres ne devraient pas être utilisées. De plus, les dysfonctionnements liés à l'utilisation de ces lois enrichissent les critiques contre le bayésien (Wilkinson, 1971) ou tout simplement contre ce type de lois. Cependant, comme nous l'avons vu précédemment, ces lois admettent un certains nombre d'avantages. Ainsi, de nombreux statisticiens tentent de faire accepter ces lois par différents moyens : soit en les faisant apparaitre comme des limites naturelles de lois *a priori* impropres, soit en revisitant les fondements des probabilités. Dans cette partie nous présentons, de façon non-exhaustive, différentes approches. Nous nous attardons particulièrement sur les différents modes de convergences considérés.

2.1 Une version relaxée de la théorie de Kolmogorov

Villegas (1967) affirme qu'une reformulation des axiomes fondateurs de la théorie des probabilités subjectives justifierait toutes les mesures *a priori*. Taraldsen and Lindqvist (2010) adhèrent à cette idée. Ils justifient l'utilisation de lois *a priori* impropres, en se basant sur la théorie des probabilités développée par Kolmogo-

rov en 1933 dont ils proposent une version relaxée. Rappelons l'axiomatique de Kolmogorov :

Définition 2.1.1. *Soit $\Omega = \{\omega_i | i \in I\}$ un ensemble appelé univers des possibles. Les ω_i sont des éventualités et une union $A = \bigcup_{j \in J} \{\omega_j\}$ est un événement. Une probabilité peut être définie sur une famille d'événements $\mathcal{A}$ si :*

- *$\mathcal{A}$ contient Ω,*
- *$\mathcal{A}$ est une σ-algèbre, c'est-à-dire que c'est une famille de sous-ensembles de Ω contenant l'ensemble vide, stable par prise du complémentaire et de l'union dénombrable.*

Définition 2.1.2. *Alors une probabilité P sur $\mathcal{A}$ est une mesure associant à chaque $A \in \mathcal{A}$ un nombre réel et vérifiant les propriétés de :*

- *Normalisation : $P(\Omega) = 1$,*
- *Positivité : $\forall A \in \mathcal{A}$, $P(A) \geqslant 0$,*
- *Additivité : si $\{A_i\}_i$ est une famille d'événements deux à deux incompatibles, alors $P(\bigcup_i A_i) = \sum_i P(A_i)$.*

Taraldsen et Lindqvist suppriment l'hypothèse de normalisation et supposent juste que la mesure est σ-finie.

Définition 2.1.3. *Soit (X, Σ, μ) un espace mesuré. On dit que la mesure μ est σ-finie lorsqu'il existe un recouvrement dénombrable de X par des sous-ensembles de mesure finie, c'est-à-dire lorsqu'il existe une suite $\{E_n\}_{n \in \mathbb{N}}$ d'éléments de la tribu Σ, tous de mesure finie, avec $X = \bigcup_{n \in \mathbb{N}} E_n$.*

Cette théorie est étroitement liée à celle des probabilités conditionnelles développée par Rényi (1970). Cependant, les motivations de Rényi (1970) n'étaient pas les mêmes : il n'avait pas pour but de développer l'inférence statistique mais avait pour intuition que les probabilités conditionnelles sont un concept fondamental. Sa théorie qui peut être considérée comme une généralisation de la théorie de Kolmogorov donne tout de même un cadre naturel pour la formulation de modèles statistiques généraux. Taraldsen and Lindqvist (2013, 2015b) appliquent leur théorie dans le cadre l'inférence fiduciaire (Fisher, 1922, 1930, 1935). Récemment, Taraldsen and Lindqvist (2015a) ont développé davantage leur théorie de 2010. Ils

définissent le concept de C-mesure comme l'espace quotient $M/\sim$ où M désigne l'ensemble des mesures σ-finies et $\sim$ est la relation d'équivalence donnée par : $\mu \sim \nu$ si et seulement si il existe $\alpha > 0$ tel que $\mu = \alpha\nu$. Ils mentionnent plusieurs fois dans leur article l'importance de l'étude d'un mode de convergence sur l'espace des C-mesures. Ceci rejoint un peu ce que nous faisons dans le chapitre 3. En effet, nous étudions un mode de convergence sur l'espace quotient $R/\sim$ où R désigne l'ensemble des mesures de Radon strictement positives et $\sim$ est la même relation d'équivalence que celle évoquée par Taraldsen and Lindqvist (2015a).

2.2 Quelques approches indirectes *via* les *a posteriori*

Une grande partie des Bayésiens sont d'accord sur le fait qu'un *a priori* impropre Π est acceptable s'il apparait comme limite d'une suite d'*a priori* propres.

Comme l'analyse bayésienne ne repose sur la distribution *a priori* qu'à travers l'*a posteriori*, certains auteurs définissent une mesure Π comme limite d'une suite $\{\Pi_n\}_n$ de façon indirecte. En effet, ils définissent des modes de convergence sur les mesures *a posteriori* et diront qu'une suite de mesures *a priori* $\{\Pi_n\}_n$ approxime l'*a priori* impropre Π si la suite de mesures *a posteriori* $\{\Pi_n(.|.)\}_n$ converge pour ce mode de convergence vers $\Pi(.|.)$. La convergence dépend donc de la vraisemblance.

Dans cette partie, nous rapportons quelques modes de convergence proposés par différents auteurs. Il sera toujours supposé que le dénominateur intervenant dans la formule de Bayes est fini, ainsi l'*a posteriori* obtenu en appliquant formellement la formule de Bayes sera toujours propre.

2.2.1 La convergence de Wallace (1959)

Wallace (1959) a montré que pour tout *a posteriori* formellement engendré par un *a priori* impropre, il existe une suite d'*a priori* propres qui fournit une suite d'*a posteriori* convergeant vers cet *a posteriori* pour chaque jeu de données fixé. Plus précisément, Wallace (1959) démontre la proposition suivante :

Proposition 2.2.1. *Si π est une densité a priori telle que $\int_\Theta \pi(\theta)\mathrm{d}\theta = +\infty$ avec pour a posteriori $\Pi(.|x)$, alors il existe une suite de densités a priori propres $\{\pi_n\}_n$ engendrant une suite d'a posteriori $\{\Pi_n(.|x)\}_n$ telle que pour tout $\theta \in \Theta$ et pour tout x,*

$$\lim_{n\to\infty} \pi_n(\theta|x) = \pi(\theta|x).$$

De plus, si $\{\pi_n\}_n$ est une suite de densités a priori telle qu'il existe une constante K et une suite $\{a_n\}_n$ telles que pour tout θ,

$$\lim_{n\to\infty} a_n\pi_n(\theta) = \pi(\theta) \tag{2.1}$$

et

$$a_n\pi_n(\theta) \leqslant K\pi(\theta) \tag{2.2}$$

alors,

$$\lim_{n\to\infty} \pi_n(\theta|x) = \pi(\theta|x). \tag{2.3}$$

La proposition 2.2.1 est une première justification en faveur de l'utilisation d'*a priori* impropres. Cependant, Stone (1965) attire l'attention sur le fait que cette approche est rétrospective puisque le jeu de données est fixé avant tout.

2.2.2 Convergence en probabilité

Stone (1965) propose une approche prospective. Il souhaite justifier l'utilisation de la mesure de Haar invariante à droite pour *a priori*. Ce travail est une généralisation de ce qu'il avait présenté dans Stone (1963, 1964). Précisons le cadre :

1. Les données x ont une distribution dépendant d'un paramètre θ. On écrira $x = (a, s)$ où a est une probabilité auxiliaire.
2. Les ensembles S des points s et Θ des points θ sont isomorphes à un groupe G de transformations g.
3. G est un groupe topologique localement compact.
4. Pour $\theta \in \Theta$, $x = (a, s)$ a une distribution de probabilité telle que la distribution de $u = \theta^{-1}s$ sachant a est indépendante de θ. De plus, la densité de probabilité de a et u par rapport à la mesure produit de λ (pour a) et μ

(pour u) existe et sera écrite $g(a,u)\mathrm{d}\lambda(a)\mathrm{d}\mu(u)$. Il n'y a pas de sous-groupe G_s de G tel que, $\int_A g(a,s)\mathrm{d}\mu(u) = 0$ pour tout ensemble A disjoint de G_s.

Stone (1965) considère les densités des mesures *a priori* par rapport à la mesure de Haar invariante à droite du groupe G. Nous noterons ν cette mesure. Il définit les densités *a priori* impropres relativement invariantes comme suit :

Définition 2.2.2. *La fonction π est une densité a priori impropre relativement invariante si $\int_\Theta \pi(\theta)\mathrm{d}\nu(\theta) = +\infty$, π est continue et $\pi(\theta_1\theta_2) = \pi(\theta_1)\pi(\theta_2)$ pour $\theta_1, \theta_2 \in \Theta$.*

Cette définition porte ici sur les densités mais coïncide avec celle d'une mesure relativement invariante. On notera $\mathcal{R}_Q$ la classe des densités *a priori* impropres relativement invariantes. Hartigan (1964) a montré que $\mathcal{R}_Q$ est une classe « naturelle » d'*a priori* a considérer car elle mène à des procédures statistiques qui sont invariantes sous les transformations qui laissent le problème invariant. Pour $v = s^{-1}\theta$, on a donc $\pi(v|x) = \pi(v|a)$.

Stone définit ensuite la convergence en probabilité. Pour cela, il introduit les suites d'*a priori* obtenues par troncature. Il considère une suite strictement croissante de compacts $\{\Theta_n\}_n$ convergeant vers Θ, puis en déduit une suite d'*a priori* $\{\Pi_n\}_n$ définie par $\pi_n(\theta) = c_n^{-1}\pi(\theta)\mathbb{1}_{\Theta_n}$ où $c_n = \int_{\Theta_n}\pi(\theta)\mathrm{d}\theta$. La définition de la convergence en probabilité est donnée en fonction de $v = s^{-1}\theta$. On note $\pi_n(v|x_n) = \pi_n(v|a_n, s_n)$ où x, a et s sont indicés pour indiquer que les données ne sont pas fixées, chaque *a priori* est évalué sur un nouveau jeu de données.

Définition 2.2.3. *La suite $\{\pi_n(v|a_n,s_n)\}_n$ converge en probabilité vers $\pi(v|a)$ si pour tout a, $\lim\limits_{n\to\infty}\pi_n(v|a,s_n)$ existe et $\lim\limits_{n\to\infty}\pi_n(v|a,s_n) = \pi(v|a)$, i.e.*

$$\forall\varepsilon > 0, \forall v, \lim_{n\to\infty}\int_{R(s_n,\varepsilon)}\pi_n(s_n|a)\mathrm{d}\mu(s_n) = 1$$

où $R(s_n,\varepsilon) = \{s_n \,/\, |\pi_n(v|a,s_n) - \pi(v|a)| < \varepsilon\}$ et $\pi(s_n|a)$ est la densité de probabilité conditionnelle de s_n dans la distribution marginale jointe de s_n, a_n obtenue en intégrant $\pi_n(\theta)g(a_n,\theta^{-1}s_n)\mathrm{d}\lambda(a_n)\mathrm{d}\mu(s_n)\mathrm{d}\nu(\theta)$.

Stone prouve qu'une condition nécessaire pour avoir la convergence en probabilité vers un *a posteriori* engendré par une densité *a priori* impropre relativement

invariante, en utilisant une suite d'*a priori* tronqués, est que la densité *a priori* impropre relativement invariante soit $\pi(\theta) = 1$. Rappelons que cette densité est donnée par rapport à la mesure de Haar invariante à droite.

Pour obtenir une condition suffisante, Stone introduit la notion de groupe Haar contrôlable.

Définition 2.2.4. *Un groupe G est dit Haar contrôlable si pour tout ensemble compact mesurable C, il existe une suite $\{G_n\}_n$ telle que $\lim_{n\to\infty} \frac{\nu(G_n[C])}{\nu(G_n)} = 1$ où $G_n[C] = \{g \,|\, gC \subset G_n\}$.*

Stone montre alors que si G est Haar contrôlable, l'*a posteriori* induit par la mesure de Haar invariante à droite comme *a priori* est limite, au sens de la convergence en probabilité, d'une suite d'*a posteriori* engendrés par des *a priori* propres.

Ainsi, Stone (1965) a adapté la définition classique de la convergence en probabilité au problème et montré que pour un tel critère, la mesure de Haar invariante à droite est un *a priori* impropre dont l'utilisation est justifiée.

Quelques années plus tard, Stone (1970) poursuit ses travaux en faveur de la mesure de Haar invariante à droite. Pour une suite $\{\pi_n\}_n$ de densités de probabilité par rapport à la mesure de Haar invariante à droite, on notera $\{\Pi_n(.|.)\}_n$ la suite d'*a posteriori* correspondants. Stone (1970) définit

$$d_n(x) = \sup_A |\Pi_n(A|x) - \Pi(A|x)|$$

pour mesurer la proximité entre $\Pi_n(.|x)$ et l'*a posteriori* invariant $\Pi(.|x)$. Si on définit $\tilde{x}_n$ et $\tilde{x}$ les variables aléatoires générées par $\pi_n(\theta)$ et $\pi(\theta^{-1}x)$ on dira que $\{\pi_n\}_n$ induit la convergence en probabilité vers $\Pi(.|.)$ si $d_n(\tilde{x}_n)$ tend vers 0 en probabilité. Stone (1970) introduit une nouvelle définition : la suite $\{\pi_n\}_n$ de densités de probabilité est dite asymptotiquement invariante à droite si

$$\lim_{n\to\infty} \int |\pi_n(\theta) - \pi_n(\theta g)| \mathrm{d}\nu(\theta) = 0$$

uniformément sur tout compact de G. Certains groupes G n'admettent pas de suite asymptotiquement invariante à droite. Alors Stone (1970) établit un théorème donnant sous certaines hypothèses une équivalence entre l'existance d'un *a*

posteriori invariant dont la convergence en probabilité est induite par une suite d'*a priori* $\{\pi_n\}_n$ et le fait que cet *a posteriori* invariant est celui induit par la mesure de Haar invariante à droite comme *a priori* et que la suite $\{\pi_n\}_n$ est asymptotiquement invariante à droite.

2.2.3 A l'aide de la distance en variation totale

Heath and Sudderth (1989) proposent une définition permettant de qualifier certains *a priori* d'approximables par des *a priori* propres. Pour cela, ils utilisent la distance en variation totale. Soient α et β deux mesures sur Θ, la distance en variation totale est définie par

$$\|\alpha - \beta\| = \sup\left(\left|\int \phi \mathrm{d}\alpha - \int \phi \mathrm{d}\beta\right| : \sup|\phi| \leqslant 1, \phi \in L_\infty(\Theta)\right)$$

Pour une inférence $\widetilde{q}$, un *a priori* π de marginale m et d'*a posteriori* q, Heath and Sudderth (1989) définissent la distance

$$d_\pi(q, \widetilde{q}) = \int \|q_x - \widetilde{q}_x\| m(\mathrm{d}x).$$

Cette distance peut être vue comme la distance moyenne entre les inférences q et $\widetilde{q}$ quand l'espérance est calculée par rapport à la marginale m associée à l'*a priori* π.

Heath and Sudderth (1989) proposent la définition suivante :

Définition 2.2.5. *Une inférence $\widetilde{q}$ est approximable par des a priori propres si*

$$\inf d_\Pi(q, \widetilde{q}) = 0$$

où l'infimum est pris sur toutes les mesures Π simplement additives sur Θ et q est l'a posteriori correspondant à l'a priori Π. Si $\widetilde{\pi}$ est un a priori impropre avec $\widetilde{q}$ pour a posteriori formel, on dit que $\widetilde{\pi}$ est approximable par des a priori propres si $\widetilde{q}$ l'est.

Heath and Sudderth (1989) conjecturent que si q est approximable par des *a priori* propres, alors il peut être approximé par troncature. Cependant, ils ne le prouvent pas.

Le critère de la définition 2.2.5 étant difficile à vérifier, Heath and Sudderth (1989) en proposent donc un autre. Pour tout $K \subset \Theta$ tel que $0 < \pi(K) < \infty$,

$$\beta(K) = \int q_x(K^c) m_K(\mathrm{d}x).$$

Soit π un *a priori* impropre, si $\inf(\beta(K) : 0 < \pi(K) < \infty) = 0$, alors π est approximable par des *a priori* propres.

2.2.4 A l'aide de la distance de Kakutani

Stein (1965) utilise la définition de la distance entre deux mesures de probabilité $m^{(1)}$ et $m^{(2)}$ introduite par Kakutani (1948)

$$\delta(m^{(1)}, m^{(2)}) = \int \left[\sqrt{\frac{\mathrm{d}m^{(1)}}{\mathrm{d}m}} - \sqrt{\frac{\mathrm{d}m^{(2)}}{\mathrm{d}m}} \right]^2 \mathrm{d}m$$

où m est une mesure par rapport à laquelle les deux mesures $m^{(1)}$ et $m^{(2)}$ sont absolument continues. Il définit la distance entre deux mesures de probabilité *a priori* $\Pi^{(1)}$ et $\Pi^{(2)}$ par

$$\delta^*(\Pi^{(1)}, \Pi^{(2)}) = \mathbb{E}_{\Pi^{(1)}} \left(\delta(\Pi^{(1)}(.|x), \Pi^{(2)}(.|x)) \right)$$

c'est-à-dire comme étant l'espérance sous $\Pi^{(1)}$ de la distance donnée par δ entre les deux *a posteriori*. Il est légitime d'étendre cette définition au cas où $\Pi^{(2)}$ n'est pas forcément une mesure de probabilité mais juste une mesure positive satisfaisant $\int_\Theta f(x|\theta)\pi^{(2)}(\theta)\mathrm{d}\theta < +\infty$. Un *a priori* impropre Π sera donc jugé acceptable s'il existe une mesure de probabilité Π^ε telle que $\delta^*(\Pi^\varepsilon, \Pi) < \varepsilon$.

2.2.5 A l'aide de la divergence de Kullback-Leibler

Berger et al. (2009) proposent une autre approche pour justifier l'utilisation de certains *a priori* impropres. Pour définir une suite approximante d'un *a priori* Π, ils considèrent une suite strictement croissante de compacts $\{\Theta_n\}_n$ convergeant vers Θ, puis construisent une suite d'*a priori* $\{\Pi_n\}_n$ définie par $\pi_n(\theta) = c_n^{-1}\pi(\theta)\mathbb{1}_{\Theta_n}$ où $c_n = \int_{\Theta_n} \pi(\theta)\mathrm{d}\theta$. La suite des densités *a posteriori* $\{\pi_n(.|x)\}_n$ ainsi construite

converge vers la densité de l'*a posteriori* formel $\pi(.|x)$ au sens de la convergence logarithmique, c'est-à-dire

$$\lim_{n\to\infty} \int_{\Theta_n} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) \mathrm{d}\theta = 0.$$

Notons que la convergence logarithmique implique la convergence L^1 dont nous rappelons la définition :

Définition 2.2.6. *Soit $\{g_n\}_n$ une suite de fonctions mesurables. La suite $\{g_n\}_n$ converge vers g au sens de la convergence L^1 si*

$$\lim_{n\to\infty} \|g_n - g\|_1 = \lim_{n\to\infty} \int |g_n(t) - g(t)| \mathrm{d}t = 0.$$

En effet, ce que Berger et al. (2009) définissent comme la convergence logarithmique correspond en fait à la convergence en entropie relative, c'est-à-dire à

$$\lim_{n\to\infty} D(\Pi_n(.|x)\|\Pi(.|x)) = 0$$

où $D(\Pi_n(.|x)\|\Pi(.|x))$ est la divergence de Kullback-Leibler, c'est-à-dire

$$D(\Pi_n(.|x)\|\Pi(.|x)) = \int_{\Theta_n} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) \mathrm{d}\theta.$$

Il est évident que la convergence en entropie relative implique la convergence L^1 étant donnée la relation

$$D(\Pi_1\|\Pi_2) \geqslant \frac{1}{2\log(2)} \|\Pi_1 - \Pi_2\|_1^2$$

dont on peut trouver la preuve dans Cover and Thomas (1991). Ainsi, la convergence logarithmique est une convergence relativement forte. Cependant, au vu de l'exemple de Fraser et al. (1985) la convergence logarithmique des *a posteriori* ne semble pas suffire pour assurer que l'*a posteriori* limite fournit des résultats cohérents.

Exemple 2.2.7. *Considérons le modèle*

$$\mathcal{M} = \{f(x|\theta) = \frac{1}{3}, x \in \{[\theta/2], 2\theta, 2\theta+1\}, \theta \in \{1, 2, \ldots\}\},$$

où $[u]$ *est la partie entière de* u*. Fraser et al. (1985) montrent que l'a priori impropre* $\pi(\theta) = 1$ *fournit un a posteriori* $\pi(\theta|x) \propto f(x|\theta)$ *fortement inconsistant. Cet a posteriori mène à un intervalle de confiance pour* θ *donné par* $\{2x, 2x+1\}$ *avec probabilité a posteriori* $2/3$ *alors que du point de vu fréquentiste la probabilité serait de* $1/3$*. Pourtant, en utilisant la suite croissante de compacts définie par* $\Theta_n = \{1, \ldots, n\}$ *et en considérant la suite des a posteriori construite selon la méthode décrite précédemment, cette suite converge logarithmiquement vers* $\pi(\theta|x)$*.*

Berger et al. (2009) considèrent donc une convergence plus forte qui n'est plus juste ponctuelle en x mais globale : une suite $\{\Pi_n(.|x)\}_n$ sera dite convergente vers $\Pi(.|x)$ au sens de cette nouvelle convergence si

$$\lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Theta_n} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) p_n(x) \mathrm{d}x = 0 \tag{2.4}$$

où $p_n(x) = \int_{\Theta_n} f(x|\theta)\pi_n(\theta)\mathrm{d}\theta$. cette notion de convergence a été introduite par Berger and Bernardo (1992). Ainsi, selon Berger et al. (2009), il est légitime d'utiliser toute distribution *a priori* Π de densité continue strictement positive, telle que $\int_\Theta f(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta < +\infty$ pour tout x, et telle qu'il existe une suite d'*a posteriori* obtenus par troncature vérifiant (2.4).

Les résultats de Berger et al. (2009) restent limités car ils ne considèrent que des suites d'*a priori* obtenus par troncature. Ils proposent deux généralisations de leurs résultats : d'une part montrer que la limite ne dépend pas de la suite de compacts utilisée, d'autre part la construction de suites approximantes autrement que par troncature. Nous nous intéresserons à cette seconde généralisation dans le chapitre 5.

2.3 Et un mode de convergence directement sur les *a priori* ?

Comme nous pouvons le remarquer, les modes de convergences proposés dans la littérature pour valider ou non l'utilisation d'un *a priori* impropre portent généralement sur les suites d'*a posteriori*. Ceci est motivé par le fait que le critère important est la cohérence de l'analyse *a posteriori* engendrée par l'*a priori* impropre avec celle qui serait engendrée par des *a priori* propres. Cependant, vérifier si l'une ou l'autre des convergences proposées dans la section précédente a lieu implique le calcul des *a posteriori* ce qui n'est pas toujours évident. De plus, tous ces modes de convergence dépendent du modèle statistique à travers la vraisemblance qui intervient dans la formule de Bayes et impacte donc la loi *a posteriori*.

Il nous semble intéressant d'étudier un mode de convergence directement sur les *a priori*. Pour cela, considérons qu'un *a priori* est une mesure de Radon strictement positive, c'est-à-dire une mesure strictement positive finie sur les compacts. La quasi-totalité des *a priori* usuels sont des mesures de Radon. Le mode de convergence usuel sur les mesures de Radon est la convergence vague dont nous rappelons la définition :

Définition 2.3.1. *Soit $\{\mu_n\}_n$ et μ des mesures de Radon. La suite $\{\mu_n\}_n$ converge vaguement vers μ si pour toute fonction h continue à support compact,* $\lim_{n\to\infty} \int h\mathrm{d}\mu_n = \int h\mathrm{d}\mu$.

Profitons-en pour rappeler aussi la définition de la convergence étroite :

Définition 2.3.2. *Soit $\{\mu_n\}_n$ et μ des mesures bornées. La suite $\{\mu_n\}_n$ converge étroitement vers μ si pour toute fonction h continue bornée,* $\lim_{n\to\infty} \int h\mathrm{d}\mu_n = \int h\mathrm{d}\mu$.

Ces deux modes de convergence sont équivalents pour les suites de mesures de probabilité.

La première idée pourrait être d'étudier la limite vague des suites de mesures de Radon. Regardons ce que l'on obtiendrait sur un exemple. Considérons la suite $\Pi_n = \mathrm{Beta}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})$, la limite des *a posteriori* après avoir observés r succès pour N expériences de Bernoulli de paramètre θ est $\Pi(\theta|x) = \mathrm{Beta}(\mathrm{r}, \mathrm{N\text{-}r})$. Pour ce modèle,

cet *a posteriori* est engendré par l'*a priori* de densité $\pi^{(1)}(\theta) = \theta^{-1}(1-\theta)^{-1}$ que l'on a donc envie de définir comme étant la limite de la suite $\{\text{Beta}(\frac{1}{\text{n}}, \frac{1}{\text{n}})\}_n$. Cependant, $\{\text{Beta}(\frac{1}{\text{n}}, \frac{1}{\text{n}})\}_n$ converge étroitement, donc vaguement, vers $\Pi^{(2)}(\theta) = \frac{1}{2}(\delta_0 + \delta_1)$.

Ainsi, considérer la limite vague d'une suite d'*a priori* $\{\Pi_n\}_n$ ne semble pas donner des résultats cohérents pour l'analyse *a posteriori*.

De plus, le but étant d'approximer une mesure impropre par une suite de mesures de probabilité, au vu de la proposition suivante il est obligatoire de faire intervenir un facteur multiplicatif.

Proposition 2.3.3. *Si la suite $\{\mu_n\}_n$ de mesures de Radon strictement positives sur l'espace localement compact E converge vaguement vers la mesure de Radon strictement positive μ, alors $\mu(E) \leqslant \liminf \mu_n(E)$.*

En effet, cette proposition implique que si une mesure de Radon strictement positive Π est la limite vague d'une suite de mesures de probabilité $\{\Pi_n\}_n$, alors $\Pi(\Theta) \leqslant 1$. Autrement dit, la limite vague d'une suite de mesures de probabilité ne peut être de masse totale supérieure à 1.

Dans le chapitre 3, nous définissons un mode de convergence applicable directement sur les suites d'*a priori* et pour lequel une suite de mesures de probabilité peut admettre une limite impropre.

Chapitre 3

Approximation d'*a priori* impropres

Ce chapitre se décompose en deux parties : la section 3.1 qui contient l'article *Approximation of improper priors* accepté par Bernoulli Journal, et la section 3.2 qui regroupe quelques résultats complémentaires qui n'apparaissent pas dans l'article.

Le but de l'article *Approximation of improper priors* est de définir un mode de convergence sur les mesures de Radon strictement positives, la quasi-totalité des mesures *a priori* usuelles étant de telles mesures.

Comme pour $\alpha > 0$, les *a priori* Π et $\alpha\Pi$ fournissent le même *a posteriori*, l'idée est née de considérer l'espace quotient $\mathcal{R}/\sim$ où $\mathcal{R}$ désigne l'ensemble des mesures de Radon strictement positives et $\sim$ est la relation d'équivalence définie par

$$\Pi \sim \Pi' \iff \exists \alpha > 0,\ \Pi' = \alpha\Pi.$$

La convergence que nous définissons comme la convergence q-vague correspond à la convergence induite par la convergence vague sur cet espace quotient qui n'est autre que l'espace projectif des mesures de Radon strictement positives. Ce mode de convergence est bien intrinsèque, il est indépendant du cadre dans lequel on utilise la suite de mesures.

L'article *Approximation of improper priors* regroupe un certain nombre de résultats sur la convergence q-vague. Nous prouvons notamment l'unicité de la li-

mite à un facteur multiplicatif près, le fait que tout *a priori* impropre peut être approximé par une suite d'*a priori* propres ou encore la conservation de la convergence q-vague en cas de reparamétrisation. Nous étudions ensuite les convergences induites sur les distributions ou estimateurs *a posteriori* lorsque l'on suppose la convergence des *a priori*. Enfin, la convergence q-vague permet d'expliquer le paradoxe de Jeffreys-Lindley qui repose en fait sur une mauvaise construction de la suite d'*a priori* considérée. Chaque terme de la suite d'*a priori* est la somme d'un poids affecté à l'hypothèse nulle et d'une densité de probabilité. La limite de cette suite est obtenue en considérant la somme des limites or ceci n'a aucun sens dans l'espace quotient.

3.1 Approximation of improper prior [1]

3.1.1 Introduction

Improper priors such as flat priors (Laplace, 1816), Jeffreys priors (Jeffreys, 1946), reference priors (Berger et al., 2009) or the Haar measures (Eaton, 1989) are often used in Bayesian analysis when no prior information is available. The posterior distribution is obtained by applying the formal Bayes rule. There are several approaches to justify the use of improper priors in statistics. Taraldsen and Lindqvist (2010) explain how the theory of conditional probability spaces developed by Rényi (1970) is related to a theory for statistics that includes improper priors. Their article is based on a generalization of Kolmogorov's theory to the σ-finite measures. They show in particular by examples that this theory is different from the alternative theory of improper priors provided by Hartigan (1983). For many authors, the inference based on an improper prior Π is legitimated as limit of inferences based on proper priors Π_n. However, there are several ways to define this limit. For example, Jeffreys (1961), Stone (1970), Bernardo and Smith (1994, Proposition 5.11), Jaynes (2003) consider the convergence, for any given observation x, of the posterior distributions $\Pi_n(\cdot|x)$ to $\Pi(\cdot|x)$ for some convergence mode such as total variation. Stone (1963) consider a convergence mode involving both the posterior distribution and the marginal distribution.

All these convergence modes are related to the statistical model through the likelihood. Moreover, there is no standard convergence mode such that a sequence Π_n of proper priors may converge to an improper prior Π independently on the statistical model. Consider, for example, a sequence of normal distributions $\mathcal{N}(0, n)$ with zero mean and variance equal to n; it is often admitted that this sequence converges to the Laplace prior since for many statistical models the Bayes estimate related to $\mathcal{N}(0, n)$ converges to the Bayes estimate for the Laplace prior. A question then arises: does the limiting behaviour of a sequence of proper priors depend on the statistical model ? Is there any intrinsic convergence mode ?

The aim of this paper is to define a convergence mode on the set of prior distributions without reference to any statistical model. In Section 3.1.2, we define

1. To appear in Bernoulli Journal

this convergence mode. We show that a sequence of vague priors is related to at most one improper prior. We also show that any improper distribution can be approximated by proper distributions and reciprocally. In Section 3.1.3, we give some conditions on the likelihood to derive convergence of posterior distributions and Bayesian estimators from the convergence of prior distributions. In Section 3.1.4, we give some examples of construction of sequences of probability measures which converge to improper priors such as the Haar measure or the Jeffreys prior. In Section 3.1.6, we give a special interest in the convergence of Beta distributions. In Section 3.1.7, we revisit the Jeffreys-Lindley paradox in the light of our convergence mode.

3.1.2 Definition, properties and examples of q-vague convergence

Let X be a random variable and assume that $X|\theta \sim P_\theta,\ \theta \in \Theta$. We assume that Θ is in $\mathbb{R}$, $\mathbb{R}^p$ with $p > 1$, or a countable set. In the Bayesian paradigm, a prior distribution Π is given on Θ. In this article, we always assume that a prior Π is a positive Radon measure, that is a positive measure which is finite on compact sets. So, a prior may be proper or improper. We denote by π the density function with respect to the Lebesgue measure in the continuous case and the counting measure in the discrete case, or more generally to some σ-finite measure. If Π is a probability measure, we can use the Bayes formula to write the posterior density:

$$\pi(\theta|x) = \frac{f(x|\theta)\ \pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta} \tag{3.1}$$

where $f(x|\theta)$ is the likelihood function.

If Π is an improper measure but $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$, we can formally apply the Bayes formula to get a posterior distribution which will be proper. Now, if we replace Π by $\alpha\Pi$, for $\alpha > 0$, we obtain the same posterior distribution. So, in this case, the posterior distribution is proper and independent of changes in the scaling of the prior. If Π is an improper measure with $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta = +\infty$, we cannot apply the Bayes formula. But in this article, we allow posterior distribution to be improper and in this case we will define it by $\pi(\theta|x) = f(x|\theta)\ \pi(\theta)$ up to within a

scalar factor.

We denote by $\mathcal{C}_K(\Theta)$ the space of real-valued continuous functions on Θ with compact support and by $\mathcal{C}_K^+(\Theta)$ the positive functions in $\mathcal{C}_K(\Theta)$. When there is no ambiguity on the space, they will be simply denoted by $\mathcal{C}_K$ or $\mathcal{C}_K^+$. We also introduce the notation $\mathcal{C}_b(\Theta)$ for the space of bounded continuous functions on Θ, and $\mathcal{C}_0(\Theta)$ for the space of continuous functions g such that for all $\varepsilon > 0$, there exists a compact $K \subset \Theta$ such that for all $\theta \in K^c$, $g(\theta) < \varepsilon$. We use the notation $\Pi(h) = \int_\Theta h \mathrm{d}\Pi$ where h is a measurable real-valued function, and $|\Pi| = \Pi(1) = \int_\Theta \mathrm{d}\Pi$, the total mass of Π.

We recall the two classic kinds of convergence of measures (Bauer, 2001). A sequence of probability measures $\{\Pi_n\}_n$ converges narrowly (also said weakly) to a probability measure Π if, for every function ϕ in $\mathcal{C}_b(\Theta)$, $\{\Pi_n(\phi)\}_n$ converges to $\Pi(\phi)$. A sequence of positive Radon measures $\{\Pi_n\}_n$ converges vaguely to a positive Radon measure Π if, for every function ϕ in $\mathcal{C}_K(\Theta)$, $\{\Pi_n(\phi)\}_n$ converges to $\Pi(\phi)$. We also recall a characterization of vague convergence for a sequence of probability measures which will be useful later in the article.

Lemma 3.1.1 (Billingsley (1986) p.393)**.** *If $\{\Pi_n\}_n$ is a sequence of probability measures and Π is a probability measure, then $\{\Pi_n\}_n$ converges vaguely to Π iff for all $g \in \mathcal{C}_0(\Theta)$, $\{\Pi_n(g)\}_n$ converges to $\Pi(g)$.*

3.1.2.1 Convergence of prior distribution sequences

In this section, we define a new convergence mode for sequences of positive Radon measures. The aim is to propose a formalization of an usual practice which consists of approximate an improper prior with a sequence of proper priors.

Definition 3.1.2. *A sequence of positive Radon measures $\{\Pi_n\}_n$ is said to converge q-vaguely to a positive Radon measure Π if there exists a sequence of positive real numbers $\{a_n\}_n$ such that $\{a_n\Pi_n\}_n$ converges vaguely to Π.*

Let us justify this definition. In Formula (3.1), if we replace Π by $\alpha\Pi$, for $\alpha > 0$, we obtain the same posterior distribution, which means that the prior distribution is defined up to within a scalar factor. So, it is natural to define the equivalence

relation $\sim$ on the space of positive Radon measures by:

$$\Pi \sim \Pi' \Longleftrightarrow \exists\ \alpha > 0 \text{ such that } \Pi = \alpha\Pi'. \tag{3.2}$$

Then, it is natural to define the quotient space of positive Radon measures by the equivalence relation $\sim$. We denote by $\overline{\Pi}$ the equivalence class of Π, that is, $\overline{\Pi} = \{\widetilde{\Pi}/\exists\alpha > 0, \widetilde{\Pi} = \alpha\Pi\}$. The q-vague convergence corresponds to the standard quotient topology on this quotient space.

Remark 3.1.3. *One referee pointed out that similar quotient spaces for σ-finite measures were considered by Taraldsen and Lindqvist (2015a) to define conditional measures.*

Proposition 3.1.4. *Let $\{\Pi_n\}_n$ and Π be positive Radon measures. The sequence $\{\Pi_n\}_n$ converges q-vaguely to Π iff $\{\overline{\Pi}_n\}_n$ converges to $\overline{\Pi}$ for the quotient topology.*

Proof.

- Direct part: Assume that $\lim_{n\to\infty} \overline{\Pi}_n = \overline{\Pi}$. The space of positive Radon measures is a metrisable space so it admits a countable neighbourhood base. Thus, there exists a decreasing sequence of open sets $\{O_i\}_{i\in\mathbb{N}}$ in the space of positive Radon measures such that for all $i \in \mathbb{N}$, $\Pi \in O_i$ and $\bigcap_{i\in\mathbb{N}} O_i = \{\Pi\}$. So, for all $i \in \mathbb{N}$, $\overline{\Pi} \in \overline{O}_i$. For any O_i, there exists N_i such that for all $n > N_i$, $\overline{\Pi}_n \in \overline{O}_i$. Without lost of generality, we can choose N_i such that $N_i > N_{i-1}$. For all n such that $N_i \leqslant n < N_{i+1}$, $\Pi_n \in \mathcal{C}(O_i)$ where $\mathcal{C}(O_i) = \{\lambda x$ with $\lambda > 0$ and $x \in O_i\}$, that is, for all n such that $N_i \leqslant n < N_{i+1}$, there exists $a_n > 0$ such that $a_n\Pi_n \in O_i$. Moreover, since $\bigcap_{i\in\mathbb{N}} O_i = \{\Pi\}$, $\lim_{n\to\infty} a_n\Pi_n = \Pi$.
- Converse part: Assume that $\{a_n\Pi_n\}_n$ converges to Π. Since the canonical mapping ϕ defined by

$$\begin{array}{rccc} \phi & : & \mathcal{R} & \to & \mathcal{R}/\sim \\ & & \Pi & \mapsto & \overline{\Pi} \end{array} \tag{3.3}$$

  where $\mathcal{R}$ is the space of positive Radon measures, is continuous, $\{\phi(a_n\Pi_n)\}$ $= \{\overline{\Pi}_n\}$ converges to $\phi(\Pi) = \overline{\Pi}$.

□

The following proposition shows that a sequence of prior measures cannot converge q-vaguely to more than one limit up to within a scalar factor.

Theorem 3.1.5. *Let $\{\Pi_n\}_n$ be a sequence of priors such that $\{\Pi_n\}_n$ converges q-vaguely to both Π_a and Π_b, then necessarily there exists $\alpha > 0$ such that $\Pi_a = \alpha\Pi_b$.*

Proof. This is a direct consequence of Proposition 3.1.36 that states that $\overline{\mathcal{R}}$ is a Hausdorff space. However, we give here a direct proof that does not involve abstract topological concept.
Assume that $\{\Pi_n\}_n$ converges q-vaguely to both Π_a and Π_b. From Definition 3.1.2, there exist two sequences of positive scalars $\{a_n\}_n$ and $\{b_n\}_n$ such that $\{a_n\Pi_n\}_n$, respectively $\{b_n\Pi_n\}_n$, converges vaguely to Π_a, respectively Π_b. We have to prove that $\Pi_b = \alpha\Pi_a$ for some positive scalar α. Since $\Pi_a \neq 0$ and $\Pi_b \neq 0$, there exist h_a and h_b in $\mathcal{C}_K^+$ such that $\Pi_a(h_a) > 0$ and $\Pi_b(h_b) > 0$. Put $h_0 = h_a + h_b$, we have $\Pi_a(h_0) > 0$ and $\Pi_b(h_0) > 0$. Moreover, $\lim_{n\to\infty} a_n\Pi_n(h_0) = \Pi_a(h_0)$ and $\lim_{n\to\infty} b_n\Pi_n(h_0)$ $= \Pi_b(h_0)$. So, there exists N such that for $n \geqslant N$, $a_n\Pi_n(h_0) > 0$ and $b_n\Pi_n(h_0) > 0$. For any h in $\mathcal{C}_K$ and $n > N$, $\lim_{n\to\infty} \frac{\Pi_n(h)}{\Pi_n(h_0)} = \lim_{n\to\infty} \frac{a_n\Pi_n(h)}{a_n\Pi_n(h_0)} = \frac{\Pi_a(h)}{\Pi_a(h_0)}$ and $\lim_{n\to\infty} \frac{\Pi_n(h)}{\Pi_n(h_0)} =$ $\lim_{n\to\infty} \frac{b_n\Pi_n(h)}{b_n\Pi_n(h_0)} = \frac{\Pi_b(h)}{\Pi_b(h_0)}$. By uniqueness of the limit in $\mathbb{R}$, $\frac{\Pi_a(h)}{\Pi_a(h_0)} = \frac{\Pi_b(h)}{\Pi_b(h_0)}$. Therefore, $\Pi_a = \frac{\Pi_a(h_0)}{\Pi_b(h_0)}\Pi_b$. The result follows. □

Theorem 3.1.6 motivates to include the improper priors in the theory since it shows these are obtained naturally from limits of proper priors. This can be compared with a completion of a metric space.

Theorem 3.1.6. *Any improper measure may be approximated by a sequence of probability measures and conversely, any proper measure may be approximated by a sequence of improper measures.*

Proof.

- Consider an improper measure Π and $\{K_n\}_n$ an increasing sequence of compacts such that $\Theta = \bigcup_n K_n$. Then $\Pi_n = \Pi\mathbb{1}_{K_n}$ is a proper measure so, $\frac{1}{|\Pi_n|}\Pi_n$ is a probability measure. Moreover, $\{\Pi_n\}_n$ converges vaguely to Π, so $\{\frac{1}{|\Pi_n|}\Pi_n\}_n$ converges q-vaguely to Π.
- Let Π be a probability measure. Consider the sequence $\Pi_n = \Pi + \alpha_n\Pi'$ where Π' is an improper measure and $\{\alpha_n\}_n$ is a decreasing sequence which converges to 0. Then, for all $n \in \mathbb{N}$, Π_n is an improper measure and $\{\Pi_n\}_n$ converges q-vaguely to Π.

□

In many statistical models, there are several parameterizations of interest. We show that the q-vague convergence is invariant by change of parameterization. Consider a new parameterization $\eta = h(\theta)$ where h is a homeomorphism. We denote by $\widetilde{\Pi}_n = \Pi_n \circ h^{-1}$ and $\widetilde{\Pi} = \Pi \circ h^{-1}$ the prior distribution on η derived from the prior distribution on θ. The following proposition establishes a link between q-vague convergence of $\{\Pi_n\}_n$ and $\{\widetilde{\Pi}_n\}_n$.

Proposition 3.1.7. *Let $\{\Pi_n\}_n$ be a sequence of priors which converges q-vaguely to Π. Let h be a homeomorphism and consider the parameterization $\eta = h(\theta)$. Then $\{\widetilde{\Pi}_n\}_n$ converges q-vaguely to $\widetilde{\Pi}$.*

Proof. From the change of variables formula, $\int g(h(\theta))\mathrm{d}\Pi_n(\theta) = \int g(\eta)\ \mathrm{d}\widetilde{\Pi}_n(\eta)$ and $\int g(h(\theta))\mathrm{d}\Pi(\theta) = \int g(\eta)\ \mathrm{d}\widetilde{\Pi}(\eta)$. Moreover, if $\{\Pi_n\}_n$ converges q-vaguely to Π, from Definition 3.1.2 there exists $\{a_n\}_n$ such that $\{a_n\Pi_n\}_n$ converges vaguely to Π. Note that for all $g \in \mathcal{C}_K$, $g \circ h \in \mathcal{C}_K$. So, for all $g \in \mathcal{C}_K$, $\lim_{n\to\infty} a_n \int g(h(\theta))\ \mathrm{d}\Pi_n(\theta) = \int g(h(\theta))\ \mathrm{d}\Pi(\theta)$, that is, $\lim_{n\to\infty} a_n \int g(\eta)\ \mathrm{d}\widetilde{\Pi}_n(\eta) = \int g(\eta)\ \mathrm{d}\widetilde{\Pi}(\eta)$. Thus $\{\widetilde{\Pi}_n\}_n$ converges q-vaguely to $\widetilde{\Pi}$. □

3.1.2.2 Convergence when approximants are probabilities

In this section, the sequence of approximants $\{\Pi_n\}_n$ is assumed to be a sequence of probability measures. Then, we can establish some links between q-vague and narrow convergence.

Indeed, if $\{\Pi_n\}_n$ is a sequence of probabilities and Θ is a compact set, q-vague convergence is equivalent to narrow convergence.

More generally, we give a necessary and sufficient condition for the narrow convergence of a sequence of probabilities which converges q-vaguely to a probability. We recall that a sequence of bounded measures $\{\Pi_n\}_n$ is said to be tight if, for each $\varepsilon > 0$, there exists a compact set K such that, for all n, $\Pi_n(K^c) < \varepsilon$.

Proposition 3.1.8. *Let $\{\Pi_n\}_n$ and Π be probability measures such that $\{\Pi_n\}_n$ converges q-vaguely to Π. Then $\{\Pi_n\}_n$ converges narrowly to Π iff $\{\Pi_n\}_n$ is tight.*

Proof. Direct part: $\{\Pi_n\}_n$ converges narrowly to Π a probability measure so $\{\Pi_n\}_n$ is tight.

Converse part: Let us show that if $\{\Pi_{n_k}\}_k$ is any subsequence of $\{\Pi_n\}_n$ which converges narrowly then $\{\Pi_{n_k}\}_k$ converges to Π. From Billingsley (1986, Theorem 25.10), there exists a subsequence $\{\Pi_{n_k}\}_k$ of $\{\Pi_n\}_n$ which converges narrowly to some probability measure, say $\widetilde{\Pi}$. Since $\{\Pi_{n_k}\}_k$ is a sequence of probabilities which converges narrowly to $\widetilde{\Pi}$, from Definition 3.1.2, $\{\Pi_{n_k}\}_k$ converges q-vaguely to $\widetilde{\Pi}$. So, from Theorem 3.1.5, there exists $\alpha > 0$ such that $\Pi = \alpha\widetilde{\Pi}$, but Π and $\widetilde{\Pi}$ are probabilities. So $\Pi = \widetilde{\Pi}$. The result follows from Billingsley (1986, Corollary p.346). □

Now, we also assume that the limiting measure Π is an improper measure. Then we can give a result about the sequence $\{a_n\}_n$ which will be useful thereafter.

Lemma 3.1.9. *Let $\{\Pi_n\}_n$ be a sequence of probability measures and $\{a_n\}_n$ a sequence of positive scalars such that $\{a_n\Pi_n\}_n$ converges vaguely to Π. If Π is improper, then necessarily $\lim_{n\to\infty} a_n = +\infty$.*

Proof. We assume that $\{a_n\Pi_n\}_n$ converges vaguely to Π so, we have $\Pi(\Theta) \leqslant \liminf_n\ a_n\ \Pi_n(\Theta)$ (Bauer, 2001, Theorem 30.3). But for all $n \in \mathbb{N}$, $\Pi_n(\Theta) = 1$ so $\Pi(\Theta) \leqslant \liminf\ a_n$. Moreover, $\Pi(\Theta) = +\infty$ so $\liminf_n\ a_n = +\infty$. The result follows. □

Lemma 3.1.10 (Lang (1977) p.38)**.** *Let E be $\mathbb{R}$, $\mathbb{R}^p$ with $p > 1$ or a countable set, for all compact $K_0 \subset \left(\bigcup_{n>0} \mathring{K}_n\right) = E$, there exists a function $h \in \mathcal{C}_K(E)$ such that $\mathbb{1}_{K_0} \leqslant h \leqslant 1$.*

When a sequence of proper priors is used to approximate an improper prior, the mass tends to concentrate outside any compact set.

Proposition 3.1.11. *Let $\{\Pi_n\}_n$ be a sequence of probability measures which converges q-vaguely to an improper prior Π. Then, for any compact K in Θ, $\lim_{n\to\infty} \Pi_n(K) = 0$, and consequently, $\lim_{n\to\infty} \Pi_n(K^c) = 1$.*

Proof. From Definition 3.1.2, there exists $\{a_n\}_n$ such that $\lim_{n\to\infty} a_n\Pi_n(h) = \Pi(h)$ for any h in $\mathcal{C}_K$. From Lemma 3.1.9, $\lim_{n\to\infty} a_n = +\infty$ whereas $\Pi(h) < +\infty$, so $\lim_{n\to\infty} \Pi_n(h) = 0$. Let K_0 be a compact set in Θ. From Lemma 3.1.10, there exists a function $h \in \mathcal{C}_K$ such that $\mathbb{1}_{K_0} \leqslant h$. So $\Pi_n(K_0) \leqslant \Pi_n(h)$ and $\lim_{n\to\infty} \Pi_n(K_0) = 0$. Since $\Pi_n(K_0) + \Pi_n(K_0^c) = 1$ for all $n \in \mathbb{N}$, thus $\lim_{n\to\infty} \Pi_n(K_0^c) = 1$. □

Many authors consider that few knowledge on the parameter is represented by priors with large variance. Here, we establish some links between the q-vague convergence of priors and the convergence of the sequence of corresponding variances.

Proposition 3.1.12. *Let $\{\Pi_n\}_n$ be a sequence of probabilities on $\mathbb{R}$ such that $\mathbb{E}_{\Pi_n}(\theta)$ is a constant. If $\{\Pi_n\}_n$ converges q-vaguely to an improper prior Π whose support is $\mathbb{R}$, then $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$.*

Proof. Since $\mathbb{E}_{\Pi_n}(\theta)$ is constant, $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$ iff $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta^2) = +\infty$. For any $r > 0$, we have $\mathbb{E}_{\Pi_n}(\theta^2) \geqslant \int_{[-r,r]^c} \theta^2 \mathrm{d}\Pi_n(\theta)$ so $\mathbb{E}_{\Pi_n}(\theta^2) \geqslant r^2\ \Pi_n([-r,r]^c)$. From Proposition 3.1.11, $\lim_{n\to\infty} \Pi_n([-r,r]^c) = 1$ and then $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta^2) \geq r^2$. Since this holds for any $r > 0$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta^2) = +\infty$. □

Corollary 3.1.13. *Let $\{\Pi_n\}_n$ be a sequence of probabilities with constant mean which approximate the Lebesgue measure $\lambda_{\mathbb{R}}$. Then, necessarily, $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$.*

However, we will see in the examples in Section 3.1.5.4.a, that when we do not assume the expectation to be constant; the variance does not necessarily diverge.

3.1.2.3 Characterization of q-vague convergence

In this section we establish several sufficient conditions for the q-vague convergence of $\{\Pi_n\}_n$ to Π through their probability density function (pdf). When Θ is continuous, then π_n and π are the standard pdf with respect to the Lebesgue measure. When Θ is discrete, then $\pi(\theta_0) = \Pi(\theta = \theta_0)$, *i.e.* π is the pdf with respect to the counting measure.

When $\Theta = \{\theta_i\}_{i\in I}$ is a discrete set with $I \subset \mathbb{N}$, we give an easy-to-check characterization of the q-vague convergence.

Proposition 3.1.14. *Let $\{\Pi_n\}_n$ and Π be priors on $\Theta = \{\theta_i\}_{i\in I}$, $I \subset \mathbb{N}$. The sequence $\{\Pi_n\}_n$ converges q-vaguely to Π iff there exists a sequence of positive real numbers $\{a_n\}_n$ such that for all $i \in I$, $\lim_{n\to\infty} a_n \pi_n(\theta_i) = \pi(\theta_i)$.*

Proof. It is a direct consequence of Definition 3.1.2 applied to the discreet case. □

Now, we consider the continuous case.

Proposition 3.1.15. *Let* $\{\Pi_n\}_n$ *and* Π *be continuous priors on* Θ *in* $\mathbb{R}$ *or* $\mathbb{R}^p$ *with* $p > 1$. *Assume that:*
1) there exists a sequence of positive real numbers $\{a_n\}_n$ *such that the sequence* $\{a_n\pi_n\}_n$ *converges pointwise to* π,
2) there exists a continuous function $g : \Theta \to \mathbb{R}^+$ *and* $N \in \mathbb{N}$ *such that for all* $n > N$ *and* $\theta \in \Theta$, $a_n\pi_n(\theta) < g(\theta)$.
Then, $\{\Pi_n\}_n$ *converges q-vaguely to* Π.

Proof. Let h be in $\mathcal{C}_K(\Theta)$. Then, $a_n\ h(\theta)\ \pi_n(\theta) \leqslant \| h \|\ g\ \mathbb{1}_K(\theta)$ where $\| h \| = \max_{\theta\in\Theta} h(\theta)$. Since $\| h \|\ g\ \mathbb{1}_K(\theta)$ is Lebesgue integrable, by dominated convergence theorem, $\lim_{n\to\infty} \int a_n\ \pi_n(\theta)\ h(\theta)\ \mathrm{d}\theta = \int \pi(\theta)h(\theta)\mathrm{d}\theta$. □

The following result will be useful to establish a result in Section 3.1.4.2.

Proposition 3.1.16. *Let* $\{\Pi_n\}_n$ *and* Π *be priors. Assume that:*
1) there exists a sequence of positive real numbers $\{a_n\}_n$ *such that the sequence* $\{a_n\pi_n\}_n$ *converges pointwise to* π,
2') for any compact set K, *there exists a scalar* M *and some* $N \in \mathbb{N}$ *such that for* $n > N$, $\sup_{\theta\in K} a_n\,\pi_n(\theta) < M$.
Then, $\{\Pi_n\}_n$ *converges q-vaguely to* Π.

Proof. The proof is similar to the proof of Proposition 3.1.15 with $a_n\ \pi_n(\theta)\ h(\theta) \leqslant M\ \sup_{\theta\in K}\ |h(\theta)|\ \mathbb{1}_K(\theta)$. □

Remark 3.1.17. *Proposition 3.1.15 and Proposition 3.1.16 hold if* $\pi(\theta)$ *is the pdf with respect to any positive Radon measure.*

3.1.3 Convergence of posterior distributions and estimators

Consider the model $X|\theta \sim P_\theta$, $\theta \in \Theta$. We denote by $f(x|\theta)$ the likelihood. The priors Π_n on Θ represent our prior knowledge. We always assume that $\int_\Theta f(x|\theta)\mathrm{d}\Pi(\theta) > 0$.

For a measure Π and a measurable function g, we define the measure $g\Pi$ by $g\Pi(f) = \Pi(gf) = \int f(\theta)g(\theta)d\Pi(\theta)$ for any f whenever the integrals are defined; $g\Pi$ is also denoted $g\,d\Pi$ or $\Pi \circ g^{-1}$ by some authors.

In this paper, we define the posterior on θ, $\Pi(\cdot|x)$, by $\pi(\theta|x) \propto f(x|\theta)\ \pi(\theta)$. Thus, the posterior $\Pi(\cdot|x)$ may be proper or improper. There are three possible cases. First, if we use a proper prior, by applying the Bayes formula, we obtain a posterior which is a probability measure. If the prior is an improper measure such that $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$, we can formally apply the Bayes rule which provides a posterior probability measure by renormalization. At last, if the prior is an improper measure such that $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta = +\infty$, the posterior is an improper measure defined by $\pi(\theta|x) = f(x|\theta)\ \pi(\theta)$ up to within a scalar factor.

In this section, we study the consequences of the q-vague convergence of $\{\Pi_n\}_n$ on the posterior analysis. In the general case where the posteriors may be proper or improper, we give a result about the q-vague convergence of posteriors $\{\Pi_n(\cdot|x)\}_n$ to $\Pi(\cdot|x)$. When posteriors are probability measures, we can establish results about the narrow convergence instead of the q-vague convergence.

Proposition 3.1.18. *Let $\{\Pi_n\}_n$ be a sequence of priors which converges q-vaguely to Π. Assume that, $\theta \longmapsto f(x|\theta)$ is a non-zero continuous function on Θ. Then $\{\Pi_n(\cdot|x)\}_n$ converges q-vaguely to $\Pi(\cdot|x)$.*

Moreover, if $\{\Pi_n(\cdot|x)\}_n$ is a tight sequence of probabilities and $\Pi(\cdot|x)$ is a probability, then $\{\Pi_n(\cdot|x)\}_n$ converges narrowly to $\Pi(\cdot|x)$.

Proof. Assume that $\{\Pi_n\}_n$ converges q-vaguely to Π. From Definition 3.1.2, there exists a sequence of positive scalars $\{a_n\}_n$ such that $\{a_n\Pi_n\}_n$ converges vaguely to Π. So, for any $h \in \mathcal{C}_K$, $\lim_{n\to\infty} a_n\Pi_n(h) = \Pi(h)$. Since $f(x|\cdot)$ is a continuous function, $f(x|\cdot)h \in \mathcal{C}_K$ and $\lim_{n\to\infty} a_n\Pi_n(f(x|\cdot)h) = \Pi(f(x|\cdot)h)$. But $\Pi_n(f(x|\cdot)h) = f(x|\cdot)\Pi_n(h)$ and $\Pi(f(x|\cdot)h) = f(x|\cdot)\Pi(h)$. So, $\{a_n f(x|\cdot)\Pi_n\}_n$ converges vaguely to $f(x|\cdot)\Pi$, or equivalently $\{f(x|\cdot)\Pi_n\}_n$ converges q-vaguely to $f(x|\cdot)\Pi$.

If $\{\Pi_n(\cdot|x)\}_n$ is a tight sequence of probabilities and $\Pi(\cdot|x)$ is a probability, the second result follows from Proposition 3.1.8. □

Remark 3.1.19. *If Θ is discrete, then $f(x|\theta)$ is necessary continuous for the discrete topology.*

The following results are based on Proposition 3.1.18 with easier-to-check assumptions.

Corollary 3.1.20. *Let $\{\Pi_n\}_n$ and Π be priors. Assume that:*
1) there exists a sequence of positive real numbers $\{a_n\}_n$ such that the sequence $\{a_n\pi_n\}_n$ converges pointwise to π,
2) $\{a_n\pi_n(\theta)\}_n$ is non-decreasing for all $\theta \in \Theta$,
3) $\theta \longmapsto f(x|\theta)$ is continuous and positive,
4) all the posteriors $\Pi_n(\cdot|x)$ and $\Pi(\cdot|x)$ are proper.
Then, $\{\Pi_n(\cdot|x)\}_n$ converges narrowly to $\Pi(\cdot|x)$.

Proof. The sequence $\{a_n\, f\, \pi_n\}_n$ is a non-decreasing sequence of non-negative functions. By monotone convergence theorem, $\lim\limits_{n\to\infty} \int a_n f(x|\theta)\ \pi_n(\theta)\ \mathrm{d}\theta = \int \lim\limits_{n\to\infty} a_n\ f(x|\theta)\ \pi_n(\theta)\ \mathrm{d}\theta = \int\ f(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta$. So, $\{a_n\Pi_n(f)\}_n$ converges to $\Pi(f) > 0$. So there exists N such that for all $n > N$, $a_n\Pi_n(f) \geqslant \frac{1}{2}\Pi(f)$. Consider $\{K_m\}_m$ an increasing sequence of compact sets such that $\bigcup K_m = \Theta$. The sequence $\{K_m^c\}_m$ decreases to $\varnothing$ so $\lim\limits_{m\to\infty} \Pi(f\mathbb{1}_{K_m^c}) = 0$. Thus, for all $\varepsilon > 0$, there exists M such that, for all $m \geqslant M$, $\Pi(f\mathbb{1}_{K_m^c}) \leqslant \varepsilon$. So, for all $n > N$, $\frac{f\,\Pi_n(K_M^c)}{\Pi_n(f)} = \frac{f\,a_n\Pi_n(K_M^c)}{a_n\Pi_n(f)} \leqslant \frac{2\,a_n\Pi_n(f\mathbb{1}_{K_M^c})}{\Pi(f)} \leqslant \frac{2\Pi(f\mathbb{1}_{K_M^c})}{\Pi(f)} \leqslant \frac{2\varepsilon}{\Pi(f)}$. The second inequality comes from assumption 3). Thus, $\{\frac{f\,\Pi_n}{\Pi_n(f)}\}_n$ is tight. The result follows from Proposition 3.1.8. □

Corollary 3.1.21. *Let $\{\Pi_n\}_n$ and Π be priors. Assume that:*
1) there exists a sequence of positive real numbers $\{a_n\}_n$ such that the sequence $\{a_n\pi_n\}_n$ converges pointwise to π,
2) there exists a continuous function $g : \Theta \to \mathbb{R}^+$ such that fg is Lebesgue integrable and for all $n \in \mathbb{N}$ and $\theta \in \Theta$, $a_n\pi_n(\theta) < g(\theta)$,
3) $\theta \longmapsto f(x|\theta)$ is continuous and positive,
4) all the posteriors $\Pi_n(\cdot|x)$ and $\Pi(\cdot|x)$ are proper.
Then, $\{\Pi_n(\cdot|x)\}_n$ converges narrowly to $\Pi(\cdot|x)$.

Proof. From Proposition 3.1.15, assumptions 1) and 2) imply that $\{\Pi_n\}_n$ converges q-vaguely to Π. From assumption 2), for all n, $a_n f(x|\theta)\pi_n(\theta) \leqslant f(x|\theta)g(\theta)$. Since fg is Lebesgue integrable, by dominated convergence theorem, $\lim\limits_{n\to\infty} \int a_n\ f(x|\theta)\ \pi_n(\theta)\ \mathrm{d}\theta = \int \lim\limits_{n\to\infty} a_n\ f(x|\theta)\ \pi_n(\theta)\ \mathrm{d}\theta = \int\ f(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta$. Thus, $\{a_n\Pi_n(f)\}_n$ converges to $\Pi(f) > 0$ so there exists N such that for all $n > N$, $a_n\Pi_n(f) \geqslant \frac{1}{2}\Pi(f)$. Consider $\{K_m\}_{m\in\mathbb{N}}$ an increasing sequence of compact sets such that $\bigcup K_m = \Theta$. The sequence $\{K_m^c\}_{m\in\mathbb{N}}$ decreases to $\varnothing$ so $\lim\limits_{m\to\infty} \lambda(fg\mathbb{1}_{K_m^c}) = 0$. Thus, for

all $\varepsilon > 0$, there exists M such that for all $m \geqslant M$, $\lambda(fg\mathbb{1}_{K_m^c}) \leqslant \varepsilon$. So, for all $n > N$, $\frac{f\, a_n \Pi_n(K_M^c)}{a_n \Pi_n(f)} \leqslant \frac{2\, a_n \Pi_n(f\mathbb{1}_{K_M^c})}{\Pi(f)} \leqslant \frac{2\lambda(fg\mathbb{1}_{K_M^c})}{\Pi(f)} \leqslant \frac{2\varepsilon}{\Pi(f)}$. Thus, $\{\Pi_n(\cdot|x)\}_n$ is a tight sequence of probabilities. The result follows from Proposition 3.1.18. □

The following result will be useful to explain the Jeffreys-Lindley paradox (see Section 3.1.7).

Corollary 3.1.22. *Consider a sequence of probabilities $\{\Pi_n\}_n$ which converges vaguely to the proper measure Π. Assume that:*
1) $\theta \longmapsto f(x|\theta)$ is continuous and non-negative,
2) $f(x|\cdot) \in \mathcal{C}_0(\Theta)$.
Then, $\{\Pi_n(\cdot|x)\}_n$ converges narrowly to $\Pi(\cdot|x)$.

Proof. Since the Π_n and Π are proper measures and $f(\cdot|\theta)$ is a pdf, $\Pi_n(\cdot|x)$ and $\Pi(\cdot|x)$ are probabilities. We assume that $\{\Pi_n\}_n$ converges vaguely, and so q-vaguely, to Π and that f satisfies 1). So, from Proposition 3.1.18, $\{\Pi_n(\cdot|x)\}_n$ converges q-vaguely to $\Pi(\cdot|x)$. From Lemma 3.1.1, $\{\Pi_n(f)\}_n$ converges to $\Pi(f)$. So, there exists N such that for $n > N$, $\Pi_n(f) > \frac{\Pi(f)}{2}$. Moreover, from assumption 2), for all $\varepsilon > 0$, there exists a compact K such that for all $\theta \in K^c$, $f(\theta|x) \leqslant \varepsilon$. Thus, for all $n > N$, $\frac{f\Pi_n(K^c)}{\Pi_n(f)} \leqslant \frac{2\Pi_n(f\mathbb{1}_{K^c})}{\Pi(f)} \leqslant \frac{2\varepsilon}{\Pi(f)}$. Thus, $\{\frac{f\,\Pi_n}{\Pi_n(f)}\}_n$ is tight. The result follows from Proposition 3.1.18. □

Now, we establish some links between the q-vague convergence of $\{\Pi_n\}_n$ and the convergence of the Bayes estimates $\mathbb{E}_{\Pi_n}(\theta|x)$.

Proposition 3.1.23. *Let $\{\Pi_n\}_n$ be a sequence of priors which converges q-vaguely to Π. Assume that:*
1) $\theta \longmapsto f(x|\theta)$ is a non-zero continuous function on Θ,
2) the family $\{\Pi_n(\cdot|x)\}_n$ is a family of probabilities uniformly integrable (Billingsley, 1968, p.32).
Then, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x) = \mathbb{E}_{\Pi}(\theta|x)$.

Proof. From Proposition 3.1.18, $\{\Pi_n(\theta|x)\}_n$ converges q-vaguely to $\Pi(\theta|x)$. For all n, $\Pi_n(\cdot|x)$ and $\Pi(\cdot|x)$ are probability measures and $\{\Pi_n(\cdot|x)\}_n$ uniformly integrable implies that $\{\Pi_n(\cdot|x)\}_n$ is tight. So, from Proposition 3.1.18, $\{\Pi_n(\theta|x)\}_n$ converges narrowly to $\Pi(\theta|x)$. The result follows from Billingsley (1968, Theorem 5.4). □

We give an other version of Proposition 3.1.23 with a more restrictive but easier-to-check condition than uniform integrability.

Corollary 3.1.24. *Let $\{\Pi_n\}_n$ be a sequence of priors which converges q-vaguely to Π. Assume that $\theta \longmapsto f(x|\theta)$ is a non-zero continuous function on Θ, and that $\{\Pi_n(\cdot|x)\}_n$ is a family of probabilities such that $\{\mathrm{Var}_{\Pi_n}(\theta|x)\}_n$ is bounded above. Then $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x) = \mathbb{E}_{\Pi}(\theta|x)$.*

Proof. This is a consequence of Billingsley (1968, p.32) and Proposition 3.1.23. □

3.1.4 Some constructions of sequences of vague priors

In this section, we give some constructions of sequences of probability measures that approximate a given improper prior such as the Haar measures or the Jeffreys prior. We have shown in the proof of Proposition 3.1.6 that any improper prior may be approximated by truncation. Here we give other constructions for the Haar measure or the Jeffreys prior.

3.1.4.1 Location and scale models

The parameter θ is said to be a location parameter if there exists a pdf g such that $f(x|\theta) = g(x-\theta)$. For instance, it is the case when $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with known σ^2. The underlying group is $(\mathbb{R}, +)$ and the Haar measure $\lambda_{\mathbb{R}}$ is improper.

Proposition 3.1.25. *Let Π be a continuous probability measure on $\mathbb{R}$. Assume that the pdf $\pi(\theta)$ of Π with respect to the Lebesgue measure $\lambda_{\mathbb{R}}$ is bounded above by a continuous and increasing function and is continuous at $\theta = 0$ with $\pi(0) > 0$. We define Π_n by $\pi_n(\theta) = \frac{1}{n}\pi(\frac{\theta}{n})$. Then, $\{\Pi_n\}_{n>0}$ converges q-vaguely to $\lambda_{\mathbb{R}}$.*

Proof. Put $\pi_n(\theta) = \frac{1}{n}\pi(\frac{\theta}{n})$. Put $a_n = n$, then $\lim_{n\to\infty} a_n\pi_n(\theta) = \lim_{n\to\infty} \pi(\frac{\theta}{n}) = \pi(0) > 0$ since π is continuous at 0. Moreover, π is bounded above by a continuous and increasing function, so there exists g such that, for all $\theta \in \mathbb{R}$ and for all $n > 0$, $\pi(\frac{\theta}{n}) \leqslant g(\frac{\theta}{n}) \leqslant g(\theta)$. The result follows from Proposition 3.1.15. □

We note that Hartigan (1996) used a dual approach. He reduced the influence of the prior by letting the conditional variance σ^2 reducing to 0. He arrived at

similar conclusions. He assumed that Π is locally uniform at 0, but it is equivalent to assuming that Π is continuous and positive at 0. We replace his condition "π tail-bounded" by the condition "π bounded".

Remark 3.1.26. *Proposition 3.1.25 holds with the assumption "π bounded" instead of "π bounded above by a continuous and increasing function".*

We now study the scale model. The strictly positive parameter σ is said to be a scale parameter if $f(x|\sigma) = \frac{1}{\sigma} g(\frac{x}{\sigma})$ where g is a pdf. If σ is a scale parameter, $\log(\sigma)$ is a location parameter for $\log(X)$. Here, the concerned group is $(\mathbb{R}^+\backslash\{0\}, \times)$ and the Haar measure $\frac{1}{\sigma}\lambda_{\mathbb{R}^+\backslash\{0\}}$ is improper. The following proposition is the equivalent of Proposition 3.1.25 for the Haar measure on $(\mathbb{R}^+\backslash\{0\}, \times)$.

Corollary 3.1.27. *Let Π be a continuous probability measure on $\mathbb{R}^+\backslash\{0\}$. Assume that the pdf $\pi(\sigma)$ of Π with respect to the Lebesgue measure $\lambda_{\mathbb{R}^+\backslash\{0\}}$ is bounded above by a continuous and increasing function and is continuous at $\sigma = 1$ with $\pi(1) > 0$. We define Π_n by $\pi_n(\sigma) = \frac{1}{n}\sigma^{\frac{1}{n}-1}\pi(\sigma^{\frac{1}{n}})$. Then, $\{\Pi_n\}_{n>0}$ converges q-vaguely to $\frac{1}{\sigma}\lambda_{\mathbb{R}^+\backslash\{0\}}$.*

Proof. Put $\theta = \log(\sigma)$. From Proposition 3.1.7, $\widetilde{\pi}(\theta) = e^{\theta}\pi(e^{\theta})$ which is bounded above by the continuous and increasing function $e^{\theta}g(e^{\theta})$. The result follows from Proposition 3.1.25. □

3.1.4.2 Jeffreys conjugate priors (JCPs)

The Jeffreys prior is one of the most popular prior when no information is available, but, in many cases, is improper. Consider that the distribution $X|\theta$ belongs to an exponential family, *i.e.* $f(x|\theta) = \exp\{\theta \cdot t(x) - \phi(\theta)\}\, h(x)$, for some functions $t(x)$, $h(x)$ and $\phi(\theta)$, and $\theta \in \Theta$, where Θ is an open set in $\mathbb{R}^p$, $p \geq 1$, such that $f(x|\theta)$ is a well-defined pdf. We assume that $\phi(\theta)$ and $I_\theta(\theta)$ are continuous. These conditions are satisfied if $t(X)$ is not concentrated on an hyperplane almost surely (Barndorff-Nielsen, 1978). Druilhet and Pommeret (2012) proposed a class of conjugate priors that aims to approximate the Jeffreys prior and that is invariant with respect to smooth reparameterization. The notion of approximation was defined only from an intuitive point of view. We can now give a more rigorous approach by using the q-vague convergence.

Denote by $\pi^J(\theta) = |I_\theta(\theta)|^{1/2}$ the pdf of the Jeffreys prior with respect to the Lebesgue measure, where θ is the natural parameter of the exponential family and $I_\theta(\theta)$ is the determinant of the Fisher information matrix. The JCPs are defined through their pdf with respect to the Lebesgue measure by

$$\pi^J_{\alpha,\beta}(\theta) \propto \exp\{\alpha.\theta - \beta\phi(\theta)\}\ |I_\theta(\theta)|^{\frac{1}{2}},$$

and for a smooth reparameterization $\theta \to \eta$ by

$$\pi^J_{\alpha,\beta}(\eta) \propto \exp\{\alpha.\theta(\eta) - \beta\phi(\theta(\eta))\}\ |I_\eta(\eta)|^{\frac{1}{2}}.$$

Proposition 3.1.28. *Let $\{(\alpha_n, \beta_n)\}_n$ be a sequence of real numbers that converges to $(0,0)$. Then, for the natural parameter θ or for any smooth reparameterization η, $\{\Pi^J_{\alpha_n,\beta_n}\}_n$ converges q-vaguely to Π^J.*

Proof. Choose $\{a_n\}_n$ such that $a_n \pi^J_{\alpha_n,\beta_n}(\theta) = \exp\{\alpha_n\,\theta - \beta_n\phi(\theta)\}\ |I_\theta(\theta)|^{\frac{1}{2}}$, which converges pointwise to $|I_\theta(\theta)|^{\frac{1}{2}}$. Put $\gamma_n = (\alpha_n, \beta_n)$ and $\psi(\theta) = (\theta, -\phi(\theta))$. We have $\gamma_n \cdot \psi(\theta) = \alpha_n\,\theta - \beta_n\phi(\theta)$. By Cauchy-Schwarz inequality, $\gamma_n \cdot \psi(\theta) \leq \|\gamma_n\|\ \|\psi(\theta)\|$. Since γ_n converges to $(0,0)$, there exists N such that, for $n > N$, $\|\gamma_n\| < 1$. Let K be a compact set in Θ, by continuity of $\psi(\theta)$, since $\phi(\theta)$ is continuous, and by continuity of $I_\theta(\theta)$, there exist M_1 and M_2 such that, for all $\theta \in K$, $\|\psi(\theta)\| < M_1$ and $|I_\theta(\theta)|^{\frac{1}{2}} < M_2$. Therefore, $a_n \pi^J_{\alpha_n,\beta_n}(\theta) \leq M_2 \exp\{M_1\}$. The result follows from Proposition 3.1.16. □

Even if we have the convergence to the Jeffreys prior, we have no guaranty that $\Pi^J_{\alpha_n,\beta_n}$ is a proper prior and there is no general result to characterize this property such as in Diaconis and Ylvisaker (1979) for usual conjugate priors. For example, consider inverse gaussian models with likelihood $f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right) \mathbb{1}_{\{x>0\}}$ where $\mu > 0$ denotes the mean parameter and $\lambda > 0$ stands for the shape parameter. Considering the parameterization $\left(\psi = \frac{1}{\mu}, \lambda\right)$, the JCPs are given by $\pi^J_{\alpha,\beta}(\psi,\lambda) \propto e^{-\frac{\lambda}{2}(\alpha_1\psi^2 - 2\beta\psi + \alpha_2)}\ \psi^{-\frac{1}{2}} \lambda^{\frac{(\beta-1)}{2}}$. Druilhet and Pommeret (2012) showed that $\pi^J_{\alpha,\beta}(\psi,\lambda)$ is proper iff $\alpha_1 > 0$, $\alpha_2 > 0$ and $-\frac{1}{2} \leqslant \beta < \sqrt{\alpha_1\alpha_2}$. So, we may consider the sequences $\alpha_{1,n} = \alpha_{2,n} = \frac{1}{n}$ and $\beta_n = \frac{1}{2n}$. By Proposition 3.1.28, $\Pi^J_{\alpha_n,\beta_n}(\psi,\lambda)$ is therefore a sequence of proper priors that converges q-vaguely to the Jeffreys prior Π^J.

Remark 3.1.29. *For any continuous function g on Θ, we can define $\pi^g_{\alpha,\beta}(\theta) \propto \exp\{\alpha.\theta - \beta\phi(\theta)\}\, g(\theta)$ and $\pi^g(\theta) = g(\theta)$. Similarly to Proposition 3.1.28, it can be shown that $\{\Pi^g_{\alpha_n,\beta_n}\}$ converges q-vaguely to Π^g.*

3.1.5 Some examples

In this section we consider some usual distributions and we look at the q-vague limiting measure.

3.1.5.1 Approximation of flat prior from uniform distributions

3.1.5.1.a The discrete case

Consider $\Theta = \mathbb{N}$, and $\Pi_n = \mathcal{U}(\{0, 1, ..., n\})$ the uniform distribution on the discrete $\{0, \ldots, n\}$. Then $\{\Pi_n\}_n$ converges q-vaguely to the counting measure.

Indeed, $\pi_n(\theta) = \frac{1}{n+1}\, \mathbb{1}_{\{0,1,\ldots,n\}}(\theta)$. Put $a_n = n+1$, then, for $\theta \in \mathbb{N}$, $\lim\limits_{n\to\infty} a_n \pi_n(\theta) = \lim\limits_{n\to\infty} \mathbb{1}_{\{0,1,\ldots,n\}}(\theta) = 1$. The result follows from Proposition 3.1.14.

3.1.5.1.b The continuous case

Let $\Theta = \mathbb{R}$, and $\Pi_n = \mathcal{U}([-n, n])$ the uniform distribution on $[-n, n]$. Then $\{\Pi_n\}_n$ converges q-vaguely to the Lebesgue measure $\lambda_{\mathbb{R}}$.

It corresponds to a location model so the result follows from Proposition 3.1.25 with $\Pi = \mathcal{U}([-1, 1])$.

3.1.5.2 Poisson distribution

Here is an example where a family of proper priors does not converge q-vaguely. Let $\Theta = \mathbb{N}$ and Π_n be the Poisson distribution with $\pi_n(\theta) = \exp(-n)\frac{n^\theta}{\theta!}$. Assume that there exists Π such that $\{\Pi_n\}_n$ converges q-vaguely to Π. Then, from Proposition 3.1.14, there exists a sequence $\{a_n\}_n$ such that for all $\theta \in \Theta$, $\lim\limits_{n\to\infty} a_n\, \pi_n(\theta) = \pi(\theta)$. Consider $\theta_0 \in \Theta$ such that $\pi(\theta_0) > 0$. There exists N such that, for all $n > N$, $\pi_n(\theta_0) > 0$. Consider $\theta > \theta_0$, for all $n > N$, $\frac{\pi_n(\theta)}{\pi_n(\theta_0)} = \frac{\theta_0!}{\theta!} n^{\theta-\theta_0}$ and $\lim\limits_{n\to\infty} \frac{\pi_n(\theta)}{\pi_n(\theta_0)} = \frac{\pi(\theta)}{\pi(\theta_0)} < +\infty$. On the other side $\lim\limits_{n\to\infty} \frac{\theta_0!}{\theta!} n^{\theta-\theta_0} = +\infty$. This is a contradiction. So, there is no prior Π such that $\{\Pi_n\}_n$ converges q-vaguely to Π.

3.1.5.3 Normal distribution

Let $\Theta = \mathbb{R}$ and $\Pi_n = \mathcal{N}(0, n)$ the normal distribution with zero mean and variance equal to n. Then $\{\Pi_n\}_n$ converges q-vaguely to the Lebesgue measure on $\mathbb{R}$.

Indeed, $\pi_n(\theta) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{\theta^2}{2n}}$ and $\pi(\theta) = 1$. Put $a_n = \sqrt{2\pi n}$, $n > 0$. Then, $\{a_n \pi_n\}_{n>0}$ converges pointwise to 1. Moreover, for all n and all θ, $a_n \pi_n(\theta) < 2$. The result follows from Proposition 3.1.15.

Remark 3.1.30. *From Theorem 3.1.5, $\{\mathcal{N}(0, n)\}_{n>0}$ cannot converge to another limiting measure than the Lebesgue measure (up to within a scalar factor).*

More generally, it can be shown that the limiting measure is the same for $\{\mathcal{N}(\mu_n, n)\}_n$ where $\{\mu_n\}_n$ is a constant or a bounded sequence. So, we consider now the case where $\lim_{n\to\infty} \mu_n = +\infty$ by taking $\mu_n = n$.

Proposition 3.1.31. *We have three cases for the convergence of $\mathcal{N}(n, \sigma_n^2)$:*

1. *If $\lim_{n\to\infty} \frac{n}{\sigma_n^2} = +\infty$, then $\{\mathcal{N}(n, \sigma_n^2)\}_n$ does not converge q-vaguely.*
2. *If $\lim_{n\to\infty} \frac{n}{\sigma_n^2} = c$ with $0 < c < \infty$, then $\{\mathcal{N}(n, \sigma_n^2)\}_n$ converges q-vaguely to $e^{c\theta} \mathrm{d}\theta$.*
3. *If $\lim_{n\to\infty} \frac{n}{\sigma_n^2} = 0$, then $\{\mathcal{N}(n, \sigma_n^2)\}_n$ converges q-vaguely to $\lambda_{\mathbb{R}}$.*

Proof. For all $n > 0$, we denote by $\Pi_n = \mathcal{N}(n, \sigma_n^2)$, and by π_n the pdf with respect to the Lebesgue measure, $\pi_n(\theta) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{(\theta-n)^2}{2\sigma_n^2})$.

1. Put $\widetilde{\pi}_n(\theta) = \exp\left(-\frac{\theta^2}{2\sigma_n^2} + \frac{\theta\, n}{\sigma_n^2}\right)$ and $\widetilde{\pi}(\theta) = e^{\frac{n^2}{2\sigma_n^2}} \pi(\theta)$. So $\{\Pi_n\}_n$ converges q-vaguely iff $\{\widetilde{\Pi}_n\}_n$ converges q-vaguely. Assume that there exists $\widetilde{\Pi}$ such that $\{\widetilde{\Pi}_n\}_n$ converges q-vaguely to $\widetilde{\Pi}$. Then, there exists a sequence $\{a_n\}_n$ such that $\{a_n \widetilde{\Pi}_n\}_n$ converges vaguely to $\widetilde{\Pi}$. Since $\widetilde{\Pi} \neq 0$, there exists an interval $[A_1, A_2]$ such that $-\infty < A_1 < A_2 < +\infty$ and $0 < \widetilde{\Pi}([A_1, A_2]) < +\infty$. Consider $[B_1, B_2]$ such that $A_2 < B_1 < B_2 < +\infty$. There exists N such that for $n > N$, $\theta \longmapsto -\frac{\theta^2}{2n} + \frac{\theta\, n}{\sigma_n^2}$ is non-decreasing. For a such n, $\widetilde{\Pi}_n([B_1, B_2]) \geqslant (B_2 - B_1) \exp(-\frac{B_1}{2\sigma_n^2} + \frac{B_1\, n}{\sigma_n^2})$ and $\widetilde{\Pi}_n([A_1, A_2]) \leqslant (A_2 - A_1) \exp(-\frac{A_2}{2\sigma_n^2} + \frac{A_2\, n}{\sigma_n^2})$. So $\frac{\widetilde{\Pi}_n([B_1,B_2])}{\widetilde{\Pi}_n([A_1,A_2])} \geqslant \frac{B_2-B_1}{A_2-A_1} \exp(C(n))$ with $C(n) = \frac{n(B_1-A_2)}{\sigma_n^2} - \frac{(B_1^2-A_2^2)}{2\sigma_n^2} \geqslant \frac{n(B_1-A_2)}{2\sigma_n^2}$. Thus, $\lim_{n\to\infty} \frac{\widetilde{\Pi}_n([B_1,B_2])}{\widetilde{\Pi}_n([A_1,A_2])} = +\infty$ but $\lim_{n\to\infty} \frac{\widetilde{\Pi}_n([B_1,B_2])}{\widetilde{\Pi}_n([A_1,A_2])} = \frac{\widetilde{\Pi}([B_1,B_2])}{\widetilde{\Pi}([A_1,A_2])} < +\infty$. So, $\{\Pi_n\}_n$ does not converge q-vaguely.

2. Put $a_n = \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{n^2}{2\sigma_n^2})$. Then $\lim_{n\to\infty} a_n\pi_n(\theta) = \lim_{n\to\infty} \exp(-\frac{\theta^2}{2\sigma_n^2} + \frac{\theta n}{\sigma_n^2}) = e^{c\theta}$. Since $\lim_{n\to\infty} \frac{n}{\sigma_n^2} = c$, there exists N such that for all $n > N$, $\frac{n}{\sigma_n^2} \in [c-\varepsilon, c+\varepsilon]$. So, for all $n > N$, $\exp(-\frac{\theta^2}{2\sigma_n^2} + \frac{\theta n}{\sigma_n^2}) \leqslant \exp((c+\varepsilon)\theta)$ which is continuous. The result follows from Proposition 3.1.15.

3. This is the same reasoning as Point 2. with $\lim_{n\to\infty} a_n\ \pi_n(\theta) = 1$ and $a_n\pi_n(\theta) \leqslant 1+\varepsilon$ for all $n > N$ and N large enough.

□

Example 3.1.32. *Assume that $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and put the prior $\Pi_n = \mathcal{N}(0, n)$ on θ. Then, $\Pi_n(\theta|x) = \mathcal{N}(\frac{nx}{\sigma^2+n}, \frac{\sigma^2 n}{\sigma^2+n})$. From Section 3.1.5.3, the two first hypotheses are satisfied and $\{\mathcal{N}(0,n)\}_n$ converges q-vaguely to the Lebesgue measure $\lambda_{\mathbb{R}}$ so here, $\Pi = \lambda_{\mathbb{R}}$. Moreover, $\theta \longmapsto f(x|\theta)$ is continuous and positive on Θ and $\Pi(\cdot|x) = \mathcal{N}(x, \sigma^2)$ is proper. So, from Theorem 3.1.20, $\{\mathcal{N}(\frac{nx}{\sigma^2+n}, \frac{\sigma^2 n}{\sigma^2+n})\}_n$ converges narrowly to $\mathcal{N}(x, \sigma^2)$.*

Example 3.1.33. *To continue Example 3.1.32, $\mathrm{Var}_{\Pi_n}(\theta|x) = \frac{\sigma^2 n}{\sigma^2+n}$ is bounded above by σ^2 and the other hypothesis of Proposition 3.1.24 have already been verified in Example 3.1.32. So, from Proposition 3.1.24, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x) = \mathbb{E}_{\Pi}(\theta)$. Indeed, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \lim_{n\to\infty} \frac{nx}{\sigma^2+n} = x = \mathbb{E}_{\Pi}(\theta)$.*

3.1.5.4 Gamma distribution

3.1.5.4.a Approximation of $\Pi = \frac{1}{\theta}\mathbb{1}_{\theta>0}\mathrm{d}\theta$

Let $\Theta = \mathbb{R}_+$ and $\Pi_n = \mathrm{Gamma}(\alpha_n, \beta_n)$ the Gamma distributions with $\lim_{n\to\infty} (\alpha_n, \beta_n) = (0,0)$. We have $\pi_n(\theta) = \frac{{\beta_n}^{\alpha_n}}{\Gamma(\alpha_n)}\ \theta^{\alpha_n - 1}\ e^{-\beta_n\theta}$. Put $a_n = \frac{\Gamma(\alpha_n)}{{\beta_n}^{\alpha_n}}$. Then $a_n\pi_n(\theta) = \theta^{\alpha_n-1}\ e^{-\beta_n\theta}$ and $\{a_n\pi_n(\theta)\}_n$ converges to $\frac{1}{\theta}$. Put $g(\theta) = \frac{1}{\theta}\ \mathbb{1}_{]0,1]}(\theta) + \mathbb{1}_{]1,+\infty[}(\theta)$. The sequence $\{\alpha_n\}_n$ goes to 0 so there exists N such that for all $n > N$, $\alpha_n < 1$. So, for $n > N$ and for $\theta > 0$, $a_n\ \pi_n(\theta) \leqslant \theta^{\alpha_n - 1} \leqslant g(\theta)$. Since g is a continuous function on $\mathbb{R}_+^*$, from Proposition 3.1.15, $\{\Pi_n\}_n$ converges q-vaguely to $\frac{1}{\theta}\,\mathrm{d}\theta$.

Recall that for $\theta \sim \mathrm{Gamma(a, b)}$, $\mathbb{E}(\theta) = \frac{a}{b}$ and $\mathrm{Var}(\theta) = \frac{a}{b^2}$. We can see below that the same convergence may be obtained with different convergences of the mean and variance.

– For $\Pi_n = \text{Gamma}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})$, we have $\mathbb{E}_{\Pi_n}(\theta) = 1$ for all n and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = \lim_{n\to\infty} n = +\infty$.
– For $\Pi_n = \text{Gamma}(\frac{1}{\mathrm{n}}, \frac{1}{\sqrt{\mathrm{n}}})$, we have $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0$ and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = 1$ for all n.
– For $\Pi_n = \text{Gamma}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}^{\frac{1}{3}}})$, we have $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \lim_{n\to\infty} n^{-\frac{2}{3}} = 0$ and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = \lim_{n\to\infty} n^{-\frac{1}{3}} = 0$.
– For $\Pi_n = \text{Gamma}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}^2})$, we have $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \lim_{n\to\infty} n = +\infty$ and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = \lim_{n\to\infty} n^3 = +\infty$.
– For $\Pi_n = \text{Gamma}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}^{\frac{2}{3}}})$, we have $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = n^{-\frac{1}{2}} = 0$ and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = \lim_{n\to\infty} n^{\frac{1}{3}} = +\infty$.

More generally, if $\liminf_n \mathbb{E}_{\Pi_n}(\theta) > 0$ then $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = +\infty$, since $\text{Var}_{\Pi_n}(\theta) = \frac{\mathbb{E}_{\Pi_n}(\theta)}{\beta_n}$ with $\lim_{n\to\infty} \beta_n = 0$.

3.1.5.4.b Approximation of $\Pi = \frac{1}{\theta} e^{-\theta} \mathbb{1}_{\theta>0} \mathrm{d}\theta$

Let us show that $\{\text{Gamma}(\alpha_\mathrm{n}, 1)\}_n$ converges q-vaguely to $\frac{1}{\theta} e^{-\theta}\ \mathbb{1}_{\theta>0}\ \mathrm{d}\theta$ when $\{\alpha_n\}_n$ goes to 0. Put $\Pi_n = \text{Gamma}(\alpha_\mathrm{n}, 1)$. Then $\pi_n(\theta) = \frac{1}{\Gamma(\alpha_n)}\ \theta^{\alpha_n - 1} e^{-\theta} \mathbb{1}_{\theta>0}$ is the pdf of Π_n. Put $a_n = \Gamma(\alpha_n)$, then $a_n \pi_n(\theta) = \theta^{\alpha_n - 1}\ e^{-\theta}\ \mathbb{1}_{\theta>0}$ converges to $\pi(\theta) = \frac{1}{\theta}\ e^{-\theta}\ \mathbb{1}_{\theta>0}$. Moreover, since $\{\alpha_n\}_n$ goes to 0, there exists N such that for $n > N$, $\alpha_n < 1$. Put $g(\theta) = \frac{1}{\theta}\ \mathbb{1}_{]0,1]}(\theta) + \mathbb{1}_{]1,+\infty[}(\theta)$. So, for $n > N$ and $\theta > 0$, $a_n \pi_n(\theta) \leqslant \theta^{\alpha_n - 1} \leqslant g(\theta)$. The function g is continuous so from Proposition 3.1.15, $\{\text{Gamma}(\alpha_\mathrm{n}, 1)\}_n$ converges q-vaguely to $\frac{1}{\theta}\ e^{-\theta}\ \mathbb{1}_{\theta>0}\ \mathrm{d}\theta$. Since $\lim_{n\to\infty} \alpha_n = 0$, we necessarily have $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = 0$ and $\lim_{n\to\infty} \text{Var}_{\Pi_n}(\theta) = 0$.

3.1.6 Convergence of Beta distributions

We now consider a more complex example which often appears in literature; see for example Tuyl et al. (2009). Let X represents the number of successes in N Bernoulli trials, and θ be the probability of a success in a single trial. Since the Beta distribution and the Binomial distribution form a conjugate pair, a common prior distribution on θ is Beta(α, α) which have mean and median equal to $\frac{1}{2}$. Three 'plausible' non-informative priors were listed by Berger (1985, p.89): the Bayes-Laplace prior Beta$(1, 1)$, the Jeffreys prior Beta$(\frac{1}{2}, \frac{1}{2})$ and the improper Haldane

prior, wrote down Beta$(0,0)$, whose density is $\pi_H(\theta) = \frac{1}{\theta(1-\theta)}$ with respect to the Lebesgue measure on $]0,1[$. If we want Beta(α,α) with large variance, necessarily α must be close to 0. Thus, we choose Beta$(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})$. The density of $\Pi_n = \text{Beta}(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})$ with respect to the Lebesgue measure on $]0;1[$ is $\pi_n(\theta) = \frac{1}{B(\frac{1}{n},\frac{1}{n})}\ \theta^{\frac{1}{n}-1}(1-\theta)^{\frac{1}{n}-1}$. As mentioned, for example, by Bernardo (1979b) or Lane and Sudderth (1983), there are two possible limiting distributions for Beta$(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})$ when n goes to $+\infty$. The first one is $\frac{1}{2}(\delta_0+\delta_1)$ which is the limiting measure given by the standard probability theory. The second one is the Haldane prior Π_H which is deduced from the posterior distributions and estimators (Lehmann and Casella, 1998). We show that it depends on the space where θ lives. Choosing $]0,1[$ or $[0,1]$ does not matter for Beta$(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})$ but it matters for the limiting distributions. We may note that the Haldane prior is a Radon measure on $]0,1[$ but not on $[0,1]$ and that $\frac{1}{2}(\delta_0+\delta_1)$ is not defined on $]0,1[$.

3.1.6.1 Convergence on $]0,1[$

In this section, we study the convergences on $]0,1[$ of $\{\text{Beta}(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})\}_{n>0}$, of the sequence of posteriors and of the sequence of estimators.

Put $a_n = B(\frac{1}{n},\frac{1}{n})$, then $a_n\ \pi_n(\theta) = \theta^{\frac{1}{n}-1}\ (1-\theta)^{\frac{1}{n}-1}$ converges to $\pi_H(\theta) = [\theta(1-\theta)]^{-1}$ and for any θ and n, $a_n\ \pi_n(\theta) < 5$. Therefore, from Theorem 3.1.15, $\{\text{Beta}(\frac{1}{\mathrm{n}},\frac{1}{\mathrm{n}})\}_{n>0}$ converges q-vaguely to Π_H.

Consider the sequence of posteriors. The sequence of priors $\{\Pi_n\}_n$ converges q-vaguely to Π_H and $\theta \longmapsto f(x|\theta)$ is continuous on Θ. Then, from Lemma 3.1.18,

- if $x=0$, $\{\Pi_n(\theta|x)\}_n$ converges q-vaguely to the improper measures with pdf $\pi(\theta) = (1-\theta)^{N-1}\ \theta^{-1}$,
- if $x=N$, $\{\Pi_n(\theta|x)\}_n$ converges q-vaguely to the improper measures with pdf $\pi(\theta) = \theta^{N-1}(1-\theta)^{-1}$
- if $0<x<N$, $\{\Pi_n(\theta|x)\}_n$ converges q-vaguely to $\Pi_H(\theta|x) = \text{Beta}(x, N-x)$.

For $0<x<N$, Beta$(x,N-x)$ is proper and $\theta \longmapsto f(x|\theta)$ is continuous and positive. So, from Theorem 3.1.20, $\{\Pi_n(\theta|x)\}_{n>0}$ converges narrowly to $\Pi_H(\theta|x)$ = Beta$(x,N-x)$.

Consider now the Bayes estimators $\mathbb{E}_{\Pi_n}(\theta|x) = \frac{1\ +\ n\ x}{2\ +\ n\ N}$ which tend to $\frac{x}{N}$. So:

- If $x=0$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x=0) = 0$ whereas $\mathbb{E}_{\Pi_H}(\theta|x=0) = \frac{1}{N}$.

– If $x = N$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x = N) = 1$ whereas $\mathbb{E}_{\Pi_H}(\theta|x = N) = +\infty$.
– If $0 < x < N$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x) = \frac{x}{N} = \mathbb{E}_{\Pi_H}(\theta|x)$.

For $x = 0$ and $x = N$, $\Pi_H(\cdot|x)$ is an improper measure. In this case, $\mathbb{E}_{\Pi_H}(\theta|x) = \int_\Theta \theta \mathrm{d}\Pi_H(\theta|x)$.

3.1.6.2 Convergence on $[0, 1]$

In this section, we study the convergences on $[0, 1]$ of $\{\text{Beta}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})\}_{n>0}$, of the sequence of posteriors and of the sequence of estimators.

For all n and for $0 < t < 1$, $\Pi_n([0, t[) + \Pi_n([t, 1 - t]) + \Pi_n(]1 - t,\ 1]) = 1$. But on $]0, 1[$, $\{\text{Beta}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})\}_{n>0}$ converges q-vaguely to the improper measure Π_H, so $\lim_{n\to\infty} \Pi_n([t, 1 - t]) = 0$. Moreover, for all n, $\Pi_n([0, t[) = \Pi_n(]1 - t,\ 1])$. Thus for all $0 < t < 1$, $\lim_{n\to\infty} \Pi_n([0, t[) = \frac{1}{2}$. From Billingsley (1986, p.192), $\{\text{Beta}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})\}_{n>0}$ converges narrowly to $\frac{1}{2}(\delta_0 + \delta_1) = \Pi_{\{0,1\}}$. By Theorem 3.1.5, $\{\text{Beta}(\frac{1}{\mathrm{n}}, \frac{1}{\mathrm{n}})\}_{n>0}$ cannot converge to an other limit such as, the Haldane measure, which is not a Radon measure on $[0, 1]$.

The limit of the posterior distributions can be deduced from the limit of the prior distributions only for $x = 0$ and $x = N$.

– If $x = 0$, $\{\Pi_n(\theta|x = 0)\}$ converges narrowly to $\Pi_{\{0,1\}}(\theta|x = 0) = \delta_0$.
– If $x = N$, $\{\Pi_n(\theta|x = N)\}$ converges narrowly to $\Pi_{\{0,1\}}(\theta|x = N) = \delta_1$.
– If $0 < x < N$, $\{\Pi_n(\theta|x)\}$ converges narrowly to $\text{Beta}(x, N - x)$ whereas the posterior $\Pi_{\{0,1\}}(\theta|x)$ does not exist.

Similarly, the limit of the estimators can be deduced from the limit of the prior distributions only for $x = 0$ and $x = N$.

– If $x = 0$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x = 0) = 0 = \mathbb{E}_{\Pi_{\{0,1\}}}(\theta|x = 0)$.
– If $x = N$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x = N) = 1 = \mathbb{E}_{\Pi_{\{0,1\}}}(\theta|x = N)$.
– If $0 < x < N$, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta|x) = \frac{x}{N}$ whereas $\mathbb{E}_{\Pi_{\{0,1\}}}(\theta|x)$ does not exist.

3.1.7 The Jeffreys-Lindley paradox

Consider the standard Gaussian model $X|\theta \sim \mathcal{N}(\theta, 1)$ and the point null hypothesis $H_0 : \theta = 0$ tested against $H_1 : \theta \neq 0$. If we use the prior $\pi(\theta) = \frac{1}{2}\mathbb{1}_{\theta=0} + \frac{1}{2}\mathbb{1}_{\theta\neq 0}$ with respect to the measure $\delta_0 + \lambda_{\mathbb{R}}$, it corresponds to the mass $\frac{1}{2}$ on H_0 and the Laplace prior on H_1. The posterior probability of H_0 is

$\Pi(\theta = 0|x) = [1 + \sqrt{2\pi}e^{x^2/2}]^{-1}$ so $\Pi(\theta = 0|x) \leqslant \left[1 + \sqrt{2\pi}\right]^{-1} \approx 0.285$ whatever the data are. An alternative is to use a sequence of proper priors $\{\Pi_n\}_n$ whose pdf are $\pi_n(\theta) = \frac{1}{2}\,\mathbb{1}_{\theta=0} + \frac{1}{2}\,\mathbb{1}_{\theta\neq 0}\,\frac{1}{\sqrt{2\pi n}}e^{-\frac{\theta^2}{2n^2}}$. With these priors, we have $\pi_n(\theta = 0|x) = \left[1 + \sqrt{\frac{1}{1+n^2}}e^{\frac{n^2x^2}{2(1+n^2)}}\right]^{-1}$ which converges to 1. This limit differs from the "non-informative" answer $\left[1 + \sqrt{2\pi}e^{x^2/2}\right]^{-1}$ and is considered as a paradox. In the light of the concept of q-vague convergence, this result is not paradoxal since, as shown in Proposition 3.1.34, the sequence of priors $\{\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n^2)\}_n$ converges vaguely to $\frac{1}{2}\delta_0$, and, the limiting posterior distribution corresponds to the posterior of the limit of the prior distributions. The following proposition generalizes this example.

Proposition 3.1.34. *Consider a partition:* $\Theta = \Theta_0 \bigcup \Theta_1$ *where* $\Theta_0 = \{\theta_0\}$. *Let* $\{\widetilde{\Pi}_n\}_n$ *be a sequence of probabilities on* Θ *which converges q-vaguely to the improper measure* $\widetilde{\Pi}$ *and such that* $\widetilde{\Pi}_n(\theta_0) = \widetilde{\Pi}(\theta_0) = 0$. *Put* $\Pi_n = \rho\delta_{\theta_0} + (1-\rho)\,\widetilde{\Pi}_n$ *where* $0 < \rho < 1$, *then* $\{\Pi_n\}_n$ *converges vaguely to* $\rho\delta_{\theta_0}$.

Moreover, assume that $\theta \longmapsto f(x|\theta)$ *is continuous and belongs to* $\mathcal{C}_0$. *Then* $\{\Pi_n(\cdot|x)\}$ *converges narrowly to* $\Pi(\cdot|x)$.

Proof. From Definition 3.1.2, there exists $\{a_n\}_n$ such that $\{a_n\widetilde{\Pi}_n\}_n$ converges vaguely to $\widetilde{\Pi}$. For $g \in \mathcal{C}_K$, $\Pi_n(g) = \rho\; g(\theta_0) + (1-\rho)\;\widetilde{\Pi}_n(g) = \rho\; g(\theta_0) + \frac{1-\rho}{a_n}$ $a_n\;\widetilde{\Pi}_n(g)$. But, $\lim_{n\to\infty} a_n\widetilde{\Pi}_n(g) = \widetilde{\Pi}(g) < \infty$. So, $\lim_{n\to\infty} \frac{1-\rho}{a_n}\; a_n\;\widetilde{\Pi}_n(g) = 0$ since, from Lemma 3.1.9, $\lim_{n\to\infty} a_n = +\infty$. Thus, $\lim_{n\to\infty} \Pi_n(g) = \rho\; g(\theta_0)$. The first result follows.

The second part is a direct consequence of Theorem 3.1.22. □

In the Proposition 3.1.34, it is assumed that $\theta \longmapsto f(x|\theta) \in \mathcal{C}_0(\Theta)$. Now, we consider the case where the limit of the likelihood $f(x|\theta)$ when θ is outside of any compact is not 0 but $f(x|\theta_0)$. In that case, the limit of the posterior probabilities is the same as the limit of the prior probabilities, as stated in the following proposition.

Proposition 3.1.35. *Consider the same notations and assumptions of Proposition 3.1.34. Moreover, assume that* $\theta \longmapsto f(x|\theta)$ *is continuous and such that for all* $\varepsilon > 0$, *there exists a compact* K *such that for all* $\theta \in K^c$, $|f(x|\theta) - f(x|\theta_0)| \leqslant \varepsilon$. *Then* $\lim_{n\to\infty} \Pi_n(\theta = \theta_0|x) = \Pi(\theta = \theta_0)$ *and* $\lim_{n\to\infty} \Pi_n(\theta \neq \theta_0|x) = \Pi(\theta \neq \theta_0)$.

Proof. By Bayes formula: $\Pi_n(\theta = \theta_0|x) = \frac{\rho f(x|\theta_0)}{\rho f(x|\theta_0)+(1-\rho)\int_\Theta f(x|\theta)\mathrm{d}\widetilde{\Pi}_n(\theta)}$. But, for all $\varepsilon > 0$, there exists a compact K such that, for all $\theta \in K^c$, $|f(x|\theta) - f(x|\theta_0)| \leqslant \varepsilon$. So $\int_\Theta f(x|\theta)\mathrm{d}\widetilde{\Pi}_n(\theta) = \int_K f(x|\theta)\mathrm{d}\widetilde{\Pi}_n(\theta) + \int_{K^c} f(x|\theta)\ \mathrm{d}\widetilde{\Pi}_n(\theta)$, where:

- $(f(x|\theta_0) - \varepsilon)\ \widetilde{\Pi}_n(K^c) \leqslant \int_{K^c} f(x|\theta)\mathrm{d}\widetilde{\Pi}_n(\theta) \leqslant (f(x|\theta_0) + \varepsilon)\ \widetilde{\Pi}_n(K^c)$. From Proposition 3.1.11, $\lim_{n\to\infty} \widetilde{\Pi}_n(K^c) = 1$. So, $\lim_{n\to\infty} \int_{K^c} f(x|\theta)\ \mathrm{d}\ \widetilde{\Pi}_n(\theta) = f(x|\theta_0)$.
- There exists $g \in \mathcal{C}_K(\Theta)$ such that $0 \leqslant g \leqslant 1$ and $g\ \mathbb{1}_K = 1$. For a such g, $\lim_{n\to\infty} \int_K f(x|\theta)\ \mathrm{d}\ \widetilde{\Pi}_n(\theta) \leqslant \lim_{n\to\infty} \frac{1}{a_n} a_n \int_\Theta g(\theta)\ f(x|\theta)\ \mathrm{d}\ \widetilde{\Pi}_n(\theta) = 0$ since $\lim_{n\to\infty} a_n \int_\Theta g(\theta)\ f(x|\theta)\ \mathrm{d}\ \widetilde{\Pi}_n(\theta) = \int_\Theta g(\theta)\ f(x|\theta)\ \mathrm{d}\ \widetilde{\Pi}(\theta) < +\infty$ and $\lim_{n\to\infty} a_n = +\infty$ from Lemma 3.1.9.

Thus, $\lim_{n\to\infty} \Pi_n(\theta = \theta_0|x) = \frac{\rho f(x|\theta_0)}{\rho f(x|\theta_0)+(1-\rho)f(x|\theta_0)} = \rho = \Pi(\theta = \theta_0)$. □

To illustrate this result in a more general case, we consider an example proposed by Dauxois et al. (2006). They consider a model choice between $\mathcal{P}(m)$ the Poisson distribution, $\mathcal{B}(N, m)$ the Binomial distribution and $\mathcal{NB}(N, m)$ the Negative Binomial distribution. These models belong to the general framework of Natural Exponentiel Families (NEFs) and are determined by their variance function $V(m) = am^2 + m$ where m is the mean parameter. Thus, a null value for a relates to the Poisson NEF, a negative one to the Binomial NEF and a positive one to the Negative Binomial NEF. The prior distribution chosen on the parameter a is Π_K defined by

$$\Pi_K(a) = \begin{cases} \frac{1}{3} \text{ if } a = 0 \\ \frac{1}{3K} \text{ if } \frac{1}{a} \in \{1, ..., K\} \\ \frac{1}{3K} \text{ if } -\frac{1}{a} \in \{n_0, ..., n_0 + K - 1\} \end{cases}$$

where K is an hyperparameter. Note that $\Pi_K(a = 0) = \Pi_K(a > 0) = \Pi_K(a < 0) = \frac{1}{3}$.

Dauxois et al. (2006) showed that the sequence of posterior distributions does not converge to δ_0 as in the previous case but $\Pi_K(a = 0|X = x)$, $\Pi_K(a > 0|X = x)$ and $\Pi_K(a < 0|X = x)$ converge to the prior probabilities $\Pi_K(a = 0)$, $\Pi_K(a > 0)$ and $\Pi_K(a < 0)$ whatever the data are when $K \to +\infty$.

Acknowledgment

The authors are grateful to Professors C. P. Robert, J. Rousseau and S. Dachian for helpful discussions. We acknowledge the comments from reviewers which resulted in an improved paper.

Appendix: Properties of the quotient space

Proposition 3.1.36. *$\overline{\mathcal{R}}$ is a Hausdorff space.*

Proof. This proof is based on two results of Bourbaki (1971).

Step 1: $\mathcal{R}$ is a topological space and $\Gamma = \{\sigma_\alpha : \Pi \longmapsto \alpha\Pi,\ \alpha \in \mathbb{R}_+^*\}$ is a homeomorphism group of $\mathcal{R}$. We consider the equivalence relation: $\Pi \sim \Pi'$ iff there exists $\alpha > 0$ such that $\Pi = \alpha\Pi'$, that is, there exists $\sigma_\alpha \in \Gamma$ such that $\Pi = \sigma_\alpha(\Pi')$. So, from Bourbaki (1971, Section I.31), $\sim$ is open.

Step 2: Let us show that $G = \{(\Pi, \alpha\Pi),\ (\Pi, \alpha\Pi) \in \mathcal{R} \times \mathcal{R}\}$ which is the graph of $\sim$ is closed. Let $\{(\Pi_n, \alpha_n\Pi_n)\}_{n>0}$ be a sequence in G such that $\lim_{n\to\infty}(\Pi_n, \alpha_n\Pi_n) = (\Pi_0, \Pi_0')$. The aim is to show that $(\Pi_0, \Pi_0') \in G$, that is, (Π_0, Π_0') takes the form $(\Pi_0, \alpha_0\Pi_0)$ where $\alpha_0\Pi_0 \neq 0$. Since $\Pi_0 \neq 0$, there exists $f_0 \in \mathcal{C}_K$ such that $\Pi_0(f_0) > 0$. Moreover, $\lim_{n\to\infty} \Pi_n(f_0) = \Pi_0(f_0)$ so there exists N such that for all $n \geqslant N$, $\Pi_n(f_0) > 0$. For all $n \geqslant N$, $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \frac{\alpha_n\Pi_n(f_0)}{\Pi_n(f_0)} = \frac{\Pi_0'(f_0)}{\Pi_0(f_0)} = \alpha_0$. Thus, for all $f \in \mathcal{C}_K$, $\lim_{n\to\infty} \alpha_n\Pi_n(f) = \alpha_0\Pi_0(f)$ and $\lim_{n\to\infty} \alpha_n\Pi_n(f) = \Pi_0'(f)$. Since $\mathcal{R}$ is a Hausdorff space, $\alpha_0\Pi_0(f) = \Pi_0'(f)$. So, the graph of $\sim$, G, is closed. The result follows from Bourbaki (1971, Section I.55).

□

3.2 Quelques résultats complémentaires

In this section we give some additional results about the q-vague convergence.

3.2.1 When densities are given with respect to a σ-finite measure

In section 3.1, we consider the density functions of measures with respect to the Lebesgue or the counting measures. So, we have established several sufficient conditions for the q-vague convergence of $\{\Pi_n\}_n$ to Π through their density functions with respect to these measures which are Radon measures.

In this section, we denote by π the density function of Π with respect to a σ-finite measure μ. We give some sufficient conditions for the q-vague convergence of $\{\Pi_n\}_n$ to Π through their density functions with respect to a σ-finite measure.

Theorem 3.2.1. *Let $\{\Pi_n\}_{n\in\mathbb{N}}$ and Π be positive Radon measures. Assume that:*
1) there exists a sequence of positive real numbers $\{a_n\}_{n\in\mathbb{N}}$ such that the sequence $\{a_n\pi_n\}_{n\in\mathbb{N}}$ converges pointwise to π,
2) there exists a function $g : \Theta \to \mathbb{R}^+$ such that, for all compact set K, $g\mathbb{1}_K$ is μ-integrable and for all $n \in \mathbb{N}$ and $\theta \in \Theta$, $a_n\pi_n(\theta) < g(\theta)$.
Then, $\{\Pi_n\}_{n\in\mathbb{N}}$ converges q-vaguely to Π.

Proof. Let h be in $\mathcal{C}_K^+(\Theta)$. Then $a_n\pi_n(\theta)h(\theta) \leqslant \| h \| \mathbb{1}_K g(\theta)$ where $\|h\| = \max_{\theta\in\Theta} h(\theta)$. Since $\| h \| \mathbb{1}_K g(\theta)$ is μ-integrable, by dominated convergence theorem,

$$\lim_{n\to\infty} \int a_n\pi_n(\theta)h(\theta)\mathrm{d}\mu(\theta) = \int \pi(\theta)h(\theta)\mathrm{d}\mu(\theta).$$

□

3.2.2 When the median is constant

In Proposition 3.1.11, we state that when a sequence of proper priors is used to approximate an improper prior, the mass tends to concentrate outside any compact set. When Θ is an interval, the following proposition gives the limiting repartition of the mass when the median is a constant.

Proposition 3.2.2. *Let $\{\Pi_n\}_{n\in\mathbb{N}}$ be a sequence of probabilities on $]a,b[$ where $-\infty \leqslant a < b \leqslant +\infty$. We assume that for all n, $\mathrm{med}(\Pi_n) = m \in]a,b[$ and that $\{\Pi_n\}_{n\in\mathbb{N}}$ converges q-vaguely to an improper Radon measure Π. Then, for any $c \in]a,b[$, $\lim_{n\to\infty} \Pi_n(]a,c[) = \frac{1}{2}$ and $\lim_{n\to\infty} \Pi_n(]c,b[) = \frac{1}{2}$.*

Proof. We only give the proof for $\Pi_n(]a,c[)$. Two cases are considered.

- Assume that $c < m$. For all n, $\Pi_n(]a,c[) + \Pi_n([c,m]) + \Pi_n(]m,b[) = 1$. But, for all n, $\Pi_n(]m,b[) \leqslant \frac{1}{2}$ and, by Proposition 3.1.11, $\lim_{n\to\infty} \Pi_n([c,m]) = 0$. So $\lim_{n\to\infty} \Pi_n(]a,c[) \geqslant \frac{1}{2}$. Moreover, for all n, $\Pi_n(]a,c[) \leqslant \Pi_n(]a,m[) \leqslant \frac{1}{2}$. So $\lim_{n\to\infty} \Pi_n(]a,c[) = \frac{1}{2}$.
- Assume that $c > m$. For all n, $\Pi_n(]a,c[) = \Pi_n(]a,m[) + \Pi_n([m,c[)$ but $\Pi_n(]a,m[) \leqslant \frac{1}{2}$ and $\lim_{n\to\infty} \Pi_n([m,c[) \leqslant \lim_{n\to\infty} \Pi_n([m,c]) = 0$ from Proposition 3.1.11. So, for all n, $\lim_{n\to\infty} \Pi_n(]a,c[) \leqslant \frac{1}{2}$. But we also have $\Pi_n(]a,c[) = \Pi_n(]a,m]) + \Pi_n(]m,c[) \geqslant \Pi_n(]a,m]) \geqslant \frac{1}{2}$.

□

Choosing c close to a or b shows that the total mass concentrate equally around a and b. Note that, in Proposition 3.2.2, we may replace $\mathrm{med}(\Pi_n) = m$ by $\mathrm{med}(\Pi_n) \in [m_1, m_2]$ with $a < m_1 < m_2 < b$.

Under the same assumptions of Proposition 3.2.2, we can easily determine the limiting value of the expectation depending on the interval Θ.

Corollary 3.2.3. *Under the same notations and assumptions of Proposition 3.2.2, we have three different cases for the limit of the expectation:*

- *If $-\infty < a$ and $b = +\infty$ then $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = +\infty$.*
- *If $a = -\infty$ and $b < +\infty$ then $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = -\infty$.*
- *If $-\infty < a < b < +\infty$ then $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \frac{a+b}{2}$.*

Proof.

- Assume that $-\infty < a$ and $b = +\infty$. For b' such that $m < b' < b$,

$$\mathbb{E}_{\Pi_n}(\theta) = \int_{]a,m[} \theta \mathrm{d}\Pi_n(\theta) + \int_{[m,b']} \theta \mathrm{d}\Pi_n(\theta) + \int_{]b',b[} \theta \mathrm{d}\Pi_n(\theta).$$

  So

$$\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) \geqslant \lim_{n\to\infty} \left(a\Pi_n(]a,m[) + m\Pi_n([m,b']) + b'\Pi_n(]b',b[)\right).$$

From Proposition 3.1.11, $\lim_{n\to\infty} \Pi_n([m, b']) = 0$. Moreover, from Proposition 3.2.2, $\lim_{n\to\infty} \Pi_n(]b', b[) = \lim_{n\to\infty} \Pi_n(]a, m[) = \frac{1}{2}$. So, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) \geqslant \frac{1}{2}(a + b')$ for all $b' > m$. The result follows.

The proof is similar for the case $a = -\infty$ and $b < +\infty$.

– Now, assume that $-\infty < a < b < +\infty$.

For $0 < \varepsilon < \frac{b-a}{2}$,

$$\mathbb{E}_{\Pi_n}(\theta) = \int_{]a,a+\varepsilon[} \theta \mathrm{d}\Pi_n(\theta) + \int_{[a+\varepsilon,b-\varepsilon]} \theta \mathrm{d}\Pi_n(\theta) + \int_{]b-\varepsilon,b[} \theta \mathrm{d}\Pi_n(\theta).$$

We have

- $a\Pi_n(]a, a + \varepsilon[) \leqslant \int_{]a,a+\varepsilon[} \theta \mathrm{d}\Pi_n(\theta) \leqslant (a + \varepsilon)\Pi_n(]a, a + \varepsilon[)$
- $(a + \varepsilon)\Pi_n([a + \varepsilon, b - \varepsilon]) \leqslant \int_{[a+\varepsilon,b-\varepsilon]} \theta \mathrm{d}\Pi_n(\theta) \leqslant (b - \varepsilon)\Pi_n([a + \varepsilon, b - \varepsilon])$
- $(b - \varepsilon)\Pi_n(]b - \varepsilon, b[) \leqslant \int_{]b-\varepsilon,b[} \theta \mathrm{d}\Pi_n(\theta) \leqslant b\Pi_n(]b - \varepsilon, b[)$

Now take the limit when n goes to infinity. From Proposition 3.1.11 for the second line and from Proposition 3.2.2 for the first and the third lines, we get after summing, $\frac{1}{2}(a + b - \varepsilon) \leqslant \lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) \leqslant \frac{1}{2}(a + b + \varepsilon)$. Since these inequalities hold for any small ε, $\lim_{n\to\infty} \mathbb{E}_{\Pi_n}(\theta) = \frac{1}{2}(a + b)$.

□

We always consider a sequence of probabilities on an interval of $\mathbb{R}$ with constant median and now we look at the limiting value of variances.

Corollary 3.2.4. *Let* $\{\Pi_n\}_{n\in\mathbb{N}}$ *be a sequence of probabilities on* $]a, +\infty[$, $]-\infty, a[$ *or* $\mathbb{R}$. *Assume that* $\mathrm{med}(\Pi_n)$ *is a constant and that* $\{\Pi_n\}_{n\in\mathbb{N}}$ *converges q-vaguely to an improper prior* Π. *Then,* $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$.

Proof. Denote $m = \mathrm{med}(\Pi_n)$ and $\mu_n = \mathbb{E}_{\Pi_n}(\theta)$. Then, we have $\mathrm{Var}_{\Pi_n}(\theta) = \mathbb{E}_{\Pi_n}((\theta - \mu_n)^2)$.

– Consider the case $\Theta =]a, +\infty[$. From Corollary 3.2.3, $\lim_{n\to\infty} \mu_n = +\infty$. So,

there exists $N \in \mathbb{N}$ such that for all $n > N$, $\mu_n > m$. Thus, for $n > N$,

$$\begin{aligned}\mathbb{E}_{\Pi_n}\left((\theta-\mu_n)^2\right) &= \int_{]a,m[}(\theta-\mu_n)^2 \mathrm{d}\Pi_n(\theta) + \int_{[m,+\infty[}(\theta-\mu_n)^2\mathrm{d}\Pi_n(\theta)\\ &\geqslant \int_{]a,m[}(\theta-\mu_n)^2\mathrm{d}\Pi_n(\theta)\\ &\geqslant \frac{1}{2}(\mu_n-m)^2.\end{aligned}$$

But $\lim_{n\to\infty}(\mu_n-m)^2 = +\infty$, so

$$\lim_{n\to\infty}\mathrm{Var}_{\Pi_n}(\theta) = +\infty.$$

The proof is quite similar in the case $\Theta =]-\infty, a[$.

– Consider now the case $\Theta = \mathbb{R}$. For any $c > |m|$, if $\mu_n < m$,

$$\begin{aligned}\mathrm{Var}_{\Pi_n}(\theta) &\geqslant \int_c^{+\infty}(\theta-\mu_n)^2\mathrm{d}\Pi_n(\theta)\\ &\geqslant \int_c^{+\infty}(\theta-m)^2\mathrm{d}\Pi_n(\theta)\\ &\geqslant \int_c^{+\infty}(c-m)^2\mathrm{d}\Pi_n(\theta) = (c-m)^2\Pi_n(]c,+\infty[).\end{aligned}$$

And, for any $c > |m|$, if $\mu_n > m$,

$$\mathrm{Var}_{\Pi_n}(\theta) \geqslant \int_{-\infty}^{-c}(c+m)^2\mathrm{d}\Pi_n(\theta) = (c+m)^2\Pi_n(]-\infty,c[).$$

Thus, in all cases,

$$\mathrm{Var}_{\Pi_n}(\theta) \geqslant \max\left\{(c+m)^2\Pi_n(]-\infty,-c[), (c-m)^2\Pi_n(]c,+\infty[)\right\}.$$

From Proposition 3.2.2,

$$\lim_n \Pi_n(]-\infty,-c[) = \lim_n \Pi_n(]c,+\infty[) = \frac{1}{2}.$$

So,

$$\lim_n \mathrm{Var}_{\Pi_n}(\theta) \geqslant \frac{1}{2}\max\left\{(c+m)^2, (c-m)^2\right\}.$$

Since this inequality holds when c goes to $+\infty$, then

$$\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty.$$

□

3.2.3 A result about variances

We now give a generalization of Corollary 3.2.4 in which we do not assume the median to be constant.

Proposition 3.2.5. *Let $\{\Pi_n\}_{n\in\mathbb{N}}$ be a sequence of probabilities on $\theta \in]a, b[$, $-\infty \leqslant a < b \leqslant +\infty$. If there exists c with $a < c < b$ such that $\lim_{n\to+\infty} \Pi_n(]a, c[) = \alpha$ for some $0 < \alpha < 1$. Then,*

- $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$ *if $a = -\infty$ or $b = +\infty$ or both.*
- $\lim_{n\to\infty} \mathrm{Var}_{\Pi_n}(\theta) = \alpha(1-\alpha)(b-a)^2$ *if $-\infty < a < b < +\infty$.*

Proof. From Proposition 3.1.11, $\lim_{n\to+\infty} \Pi_n(]a, c[) = \alpha$ for some $c \in]a, b[$ is equivalent to $\lim_{n\to+\infty} \Pi_n(]a, a'[) = \alpha$ for any $a' \in]a, b[$ which is also equivalent to $\lim_{n\to+\infty} \Pi_n(]b', b[) = 1 - \alpha$ for any $b' \in]a, b[$.

Step 1: For all $n \in \mathbb{N}$,

$$\mathrm{Var}_{\Pi_n}(\theta) = \frac{1}{2} \int\int (x-y)^2 \mathrm{d}\Pi_n(x)\mathrm{d}\Pi_n(y).$$

So, for any $a < a' < b' < b$,

$$\begin{aligned}
\mathrm{Var}_{\Pi_n}(\theta) &\geqslant \int\int_{]a,a'[\times]b',b[} (x-y)^2 \mathrm{d}\Pi_n(x)\mathrm{d}\Pi_n(y) \\
&\geqslant (b'-a')^2 \int\int_{]a,a'[\times]b',b[} \mathrm{d}\Pi_n(x)\mathrm{d}\Pi_n(y) \\
&\geqslant (b'-a')^2 \Pi_n(]a, a'[)\Pi_n(]b', b[).
\end{aligned}$$

So $\lim_{n\to+\infty} \mathrm{Var}_{\Pi_n}(\theta) \geqslant (b'-a')^2\alpha(1-\alpha)$ for all a', b' such that $a < a' < b' < b$. Taking $a' \to a$ and $b' \to b$, we get $\lim_{n\to+\infty} \mathrm{Var}_{\Pi_n}(\theta) \geqslant (b-a)^2\alpha(1-\alpha)$ if $-\infty < a < b < +\infty$ and $\lim_{n\to+\infty} \mathrm{Var}_{\Pi_n}(\theta) = +\infty$ if $a = -\infty$ or $b = +\infty$.

Step 2: For any $a < a' < b' < b$, we denote by $A_1 =]a, a'[$, $A_2 = [a', b']$, $A_3 =]b', b[$ and $B_{ij} = A_i \times A_j$, $(i, j) \in \{1, 2, 3\}^2$. For all $n \in \mathbb{N}$,

$$\mathrm{Var}_{\Pi_n}(\theta) = \sum_{i,j} \int \int_{B_{ij}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y).$$

We have the following inequalities:

- $\int \int_{B_{11}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant (a - a')^2 \Pi_n(B_{11})$
- $\int \int_{B_{22}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant (b' - a')^2 \Pi_n(B_{22})$
- $\int \int_{B_{33}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant (b - b')^2 \Pi_n(B_{11})$
- $\int \int_{B_{12} \bigcup B_{21}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant 2(b' - a)^2 \Pi_n(B_{12})$
- $\int \int_{B_{32} \bigcup B_{23}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant 2(b - a')^2 \Pi_n(B_{23})$
- $\int \int_{B_{31} \bigcup B_{13}} (x - y)^2 \mathrm{d}\Pi_n(x) \mathrm{d}\Pi_n(y) \leqslant 2(b - a)^2 \Pi_n(B_{23})$

And,

- $\lim\limits_{n \to \infty} \Pi_n(B_{11}) = \Pi_n(A_1) \times \Pi_n(A_1) = \alpha^2$,
- $\lim\limits_{n \to \infty} \Pi_n(B_{22}) = 0$,
- $\lim\limits_{n \to \infty} \Pi_n(B_{33}) = (1 - \alpha)^2$,
- $\lim\limits_{n \to \infty} \Pi_n(B_{12}) = 0$,
- $\lim\limits_{n \to \infty} \Pi_n(B_{23}) = 0$,
- $\lim\limits_{n \to \infty} \Pi_n(B_{13}) = \alpha(1 - \alpha)$.

So,

$$\lim_{n \to +\infty} \mathrm{Var}_{\Pi_n}(\theta) \leqslant \alpha^2 (a - a')^2 + (b - b')^2 (1 - \alpha)^2 + (b - a)^2 \alpha (1 - \alpha).$$

When a' tends to a and b' tends to b, we have $\lim\limits_{n \to +\infty} \mathrm{Var}_{\Pi_n}(\theta) \leqslant \alpha\ (1 - \alpha)\ (b - a)^2$. Combining with Step 1, we get $\lim\limits_{n \to +\infty} \mathrm{Var}_{\Pi_n}(\theta) = \alpha\ (1 - \alpha)\ (b - a)^2$ if $-\infty < a < b < +\infty$.

□

Chapitre 4

Utilisation de lois vagues en Removal Sampling

Dans ce chapitre, nous utilisons les résultats obtenus grâce à la convergence q-vague pour fournir des recommandations sur le choix des *a priori* dans le cadre du removal sampling. Dans la section 4.1 , nous présentons la méthode de removal sampling ainsi que les différentes méthodes utilisées pour l'estimation des paramètres. La section 4.2 contient un article dans lequel nous étudions de manière théorique les propriétés du modèle associé au removal sampling. Nous avons également mené des simulations afin d'une part d'illustrer les résultats théoriques et d'autre part de fournir des conseils aux utilisateurs.

4.1 Introduction

4.1.1 La méthode de removal sampling

L'échantillonnage par prélèvements successifs ou « removal sampling » en anglais consiste à répéter des échantillonnages sur une même unité d'observation. Une unité d'observation est une zone échantillonnée par removal sampling. Les individus sont capturés successivement et sans remise parmi la population (Williams et al., 2002b). L'intervalle de temps entre deux prélèvements successifs est généralement court pour préserver l'hypothèse majeure de population fermée, c'est-à-dire

pas d'immigration, de naissance ou de mort (MacKenzie and Royle, 2005). Les quantités successives d'individus capturés sont ensuite modélisées pour estimer la probabilité de détection et la taille de la population présente sur l'unité d'observation.

Le vecteur $X = (X_1, X_2, \ldots, X_K)$ représente la séquence de K captures observées sur une unité d'observation donnée. On pose X_k le nombre de captures au $k^{\text{ème}}$ échantillonnage, N_k la taille de la population restante après le $k^{\text{ème}}$ échantillonnage avec $N_k = N_{k-1} - X_k$ pour $k \in \{1, \ldots, K\}$ et τ_k le taux d'échantillonnage au rang k. On considère alors que les X_k suivent une distribution binomiale de paramètre N_{k-1} et τ_k :

$$(X_k | N_{k-1}, \tau_k) \sim \text{Bin}(N_{k-1}, \tau_k).$$

La probabilité de capture τ_k représente la probabilité pour un individu d'être détecté au $k^{\text{ème}}$ échantillonnage. Le taux d'échantillonnage représente la proportion de la population détectée par la technique d'échantillonnage. On suppose que les individus de la population N_k sont capturés indépendamment les uns des autres et avec la même probabilité τ_k ; le taux d'échantillonnage est alors égal à la probabilité de capture.

Le taux d'échantillonnage peut être envisagé comme plus ou moins constant selon les conditions liées aux caractéristiques de l'espèce étudiée, de l'unité d'observation ou des propriétés connues ou non de la technique d'échantillonnage. Différents modèles peuvent alors être envisagés, allant du plus simple (taux constant) vers des modèles plus complexes (taux variable).

Dans la section 4.2, le taux d'échantillonnage sera supposé constant au cours des échantillonnages successifs (Moran, 1951; Zippin, 1958; Dodd and Dorazio, 2004; Royle, 2004b,a; Dorazio et al., 2006), c'est-à-dire qu'on aura :

$$\tau_k = \tau,\ k = 1, \ldots, K.$$

Il est aussi possible de considérer un modèle à taux variable, c'est-à-dire un modèle dont le taux d'échantillonnage varie au cours des K échantillonnages successifs :

$$\tau_k \neq \tau_{k'},\ \text{pour}\ k \neq k'.$$

Cependant, les τ_k ne sont pas identifiables car en considérant une seule unité d'observation, nous ne disposons que d'une seule observation X_k par rang d'échantillonnage k. Ce problème d'identifiabilité peut être contourné en définissant une structure temporelle particulière sur le taux déchantillonnage ; par exemple, le taux d'échantillonnage de rang k peut être défini comme égal à celui du rang $k-1$ à un facteur aléatoire près, indépendant de k noté ε (Dauphin et al., 2009; Brun et al., 2011) :

$$\text{logit}(\tau_k) = \text{logit}(\tau_{k-1}) - \varepsilon \text{ où } \text{logit}(\tau) = \ln\left(\frac{\tau}{1-\tau}\right).$$

Cette structure temporelle annule le problème d'identifiabilité en réduisant le nombre de paramètres à estimer : deux paramètres (τ_1 et ε) contre K paramètres $(\tau_1, \ldots, \tau_K)$.

D'autres modélisations peuvent être envisagées dans le cas où on considère simultanément m unités d'observation échantillonnées par removal sampling. On note alors N_0^i la taille de la population sur l'unité d'observation i. On pose $X^i = (X_1^i, \ldots, X_K^i)$ la séquence de captures observées sur l'unité d'observation i pour $i = 1, \ldots, m$ où K représente le nombre total d'échantillonnages successifs réalisés sur chaque unité d'observation. On note X_k^i le nombre de captures au $k^{\text{ème}}$ échantillonnage pour l'unité d'observation i, N_k^i la taille de la population restante après le $k^{\text{ème}}$ échantillonnage et τ_k^i le taux d'échantillonnage au rang k pour le site i. On considère que les X_k^i suivent une distribution binomiale de paramètres N_{k-1}^i et τ_k^i :

$$(X_k^i | N_{k-1}^i, \tau_k^i) \sim \text{Bin}(N_{k-1}^i, \tau_k^i).$$

Ce modèle n'est pas identifiable pour tous les τ_k^i et N_0^i, il faut donc définir une structure sur les paramètres.

4.1.2 Estimation des paramètres

Comme toujours, deux grands types d'approches sont proposés pour estimer la taille de la population N_0 et le taux d'échantillonnage τ à partir des données de captures obtenues par removal sampling : l'approche fréquentiste et l'approche bayésienne. Nous présentons ici les principaux éléments.

4.1.2.1 Approche fréquentiste

La méthode des moindres carrés utilisée par Leslie and Davis (1939) et Hayne (1949) consiste à estimer le taux d'échantillonnage par la pente de la droite de régression et la taille de la population par le point d'intersection de la droite de régression avec l'axe des abscisses. Cette méthode est facile à mettre en oeuvre. Cependant, elle est connue pour fournir de mauvaises estimations de N_0 en estimant parfois une taille de population inférieure au total des captures (Schnute, 1983).

Une approche par maximum de vraisemblance peut également être considérée. Moran (1951) propose un modèle multinomial pour estimer simultanément le couple (N_0, τ) à partir des propiétés asymptotiques de l'estimateur du maximum de vraisemblance. Zippin (1956, 1958) propose une méthode graphique basée sur le maximum de vraisemblance pour estimer les paramètres. Une approche par maximum de vraisemblance permet une modélisation plus riche que la méthode des moindres carrés. Cependant, cette méthode présente de nombreuses limites dans le cadre du removal sampling :

- L'approche par maximum de vraisemblance conduit régulièrement à des estimations infinies pour N_0 (Carle and Strub, 1978; Schnute, 1983; Bolfarine et al., 1992; Bedrick, 1994).
- Lorsque le maximum de vraisemblance converge, les estimations privilégiées de N_0 sont des estimations « basses » qui sont le plus souvent très proches de la somme cumulée des captures (Schnute, 1983). Plusieurs auteurs (Schnute, 1983; Gove et al., 1995) associent ce « biais conditionnel » de l'estimateur de N_0 au fait que l'hypothèse de taux d'échantillonnage constant n'est pas valide.
- Les intervalles de confiance des estimateurs sont souvent basés sur l'approximation normale mais cette approximation n'est valable que pour des N_0 « grands ». Ces intervalles de confiance sont alors peu fiables et comprennent souvent des valeurs aberrantes (borne inférieure de l'intervalle inférieure à la somme cumulée des captures). Hirst (1994) propose des intervalles de confiance basés sur les rapports de log-vraisemblances profilées. Il démontre par simulations que ces intervalles de confiance sont plus proches de la vraie

valeur que ceux basés sur la vraisemblance asymptotique. Cependant, ces approximations asymptotiques ne sont pas valables lorsque le taux d'échantillonnage est faible (Carle and Strub, 1978).

4.1.2.2 Approche Bayésienne

Compte tenu des limites des méthodes fréquentistes, de nombreux auteurs se sont tournés vers des approches bayésiennes (Bolfarine et al., 1992; Ellison, 2004; Schwarz and Seber, 1999). D'un point de vue pratique, le choix de la loi *a priori* est souvent délicat car la connaissance *a priori* disponible est le plus souvent insuffisante pour permettre de déterminer une loi *a priori* précise. Pour le taux d'échantillonnage, la distribution intuitive est une loi uniforme sur $[0, 1]$ (Dorazio et al., 2006; Laplace, 1786) qui correspond à une loi Beta$(1, 1)$. Certains auteurs proposent l'*a priori* de Haldane (1932) qui correspond à une loi Beta$(0, 0)$ qui équivaut à une loi uniforme sur le logit, ou encore l'*a priori* de Jeffreys (1946) qui correspond à une loi Beta$(1/2, 1/2)$ qui équivaut à une loi uniforme sur $\sin^{-1}(\sqrt{\theta})$. Pour la taille de la population, en cas d'absence totale d'information, la distribution *a priori* la plus intuitive est une loi uniforme sur $\mathbb{N}$.

L'approche bayésienne est aussi utilisée dans le cadre de modèles plus évolués, notamment dans le cas où on considère plusieurs unités d'observation.

Dauphin et al. (2009) considèrent le cas où le taux d'échantillonnage est variable selon les rangs d'échantillonnage k mais constant selon les unités d'observations pour une même valeur de k. Ils considèrent un effet alétaoire ε sur lequel ils posent un *a priori* vague $\varepsilon \sim \mathcal{N}(0, 1000)$.

Dans le cas où τ^i est un effet alétaoire avec $\tau^i \sim \text{Beta}(a, b)$, Bohrmann et al. (2012) considèrent des lois *a priori* vagues Gamma$(0.01, 0.01)$ sur a et b.

Rivot et al. (2008) considèrent τ comme un effet aléatoie avec $\text{logit}(\tau^i) \sim \mathcal{N}(\mu, \sigma^2)$ où μ est distribué selon une loi *a priori* $\mathcal{N}(0, 1000)$ et σ suit une loi *a priori* $\mathcal{U}([0, 1])$.

Enfin, dans le cas où le taux d'échantillonnage est variable entre les unités d'observation et selon le rang d'échantillonnage, le modèle n'est pas identifiable. Cependant, Mantyniemi et al. (2005) contournent le problème en définissant une structure temporelle et spatiale du taux où $\tau_k^i = \mu^i \frac{\eta^i}{\eta^i + k - 1}$. Ils posent une loi *a priori*

Beta$(1.1, 1.1)$ sur μ^i et $\mathcal{U}([0,1])$ sur η^i avec i le site et k le rang d'échantillonnage.

4.1.3 Choix d'*a priori* en removal sampling

Dans la plupart des modèles statistiques, la vraisemblance tend vers 0 sur les bords du domaine étudié. Ceci entraîne la convergence de l'estimateur du maximum de vraisemblance ainsi qu'une certaine stabilité des estimateurs bayésiens avec les *a priori* vagues. Ce n'est pas le cas pour le removal sampling.

Dans l'article *Bayesian estimation of abundance by removal sampling*, nous nous intéressons à la convergence et la stabilité des estimateurs bayésiens. Nous montrons que le modèle removal sampling a pour modèle limite un modèle de Poisson iid quand N_0 tend vers l'infini, τ tend vers 0, et $N_0\tau$ vers une constante strictement positive. Ce modèle limite n'est pas identifiable ce qui est problématique pour l'estimation des différents paramètres. L'impact de ce phénomène sur l'analyse bayésienne est important. En effet, si le poids de l'*a priori* est trop fort sur des valeurs de τ faibles et/ou des valeurs de N_0 élevées, alors le fait que la vraisemblance du modèle ne tende pas vers 0 peut mener à des *a posteriori* impropres ou des estimateurs divergents ou instables. Dans l'article, nous établissons des conditions nécessaires et suffisantes sur les *a priori* pour obtenir des *a posteriori* propres et des estimateurs convergents. Puis, à l'aide de la convergence q-vague, nous montrons que les estimateurs obtenus avec des *a priori* vagues sont très instables. En effet, ils montrent une grande dépendances aux hyperparamètres. Ainsi, nous mettons en garde les utilisateurs qui travaillent avec des *a priori* vagues pour approcher un *a priori* impropre, puis considèrent un estimateur obtenu par passage à la limite sur les hyperparamètes.

4.2 Bayesian estimation of abundance by removal sampling [1]

4.2.1 Introduction

The removal method is commonly used in ecology to estimate the abundance of animal populations (Seber, 1982; Williams et al., 2002a). This sampling method is widely applied in fishery abundance studies (Wyatt, 2002; Mantyniemi et al., 2005; Dorazio and Jelks, 2005; Royle and Dorazio, 2006; Dauphin et al., 2009; Brun et al., 2011) but has also been used in studies of amphibian (Heyer et al., 1994; Bailey et al., 2004) and ticks (Bord et al., 2014).

Removal sampling consists of capturing individuals over successive samplings occasions in a single point of observation. The captured individuals are removed from the population. At each sampling, each individual in the population is assumed to have the same probability of capture, which may vary with the rank of sampling. For example, Dauphin et al. (2009) and Brun et al. (2011) considered the sampling rate to have a temporal structure, which varied by a random value ϵ. When successive samplings are conducted over a short period of time, it is common to assume a closed population and constant probability of capture over samplings.

To estimate the population size N_0 and the probability of capture τ based on removal sampling data $(X_1, \ldots, X_k)$, the asymptotic maximum likelihood approach has been used by many authors (Moran, 1951; Zippin, 1956; Seber, 1982). However, estimations based on the likelihood function may fail for several reasons. Firstly, the likelihood function may return to infinite estimates of population size N_0 with a non null-probability. Bedrick (1994) gives a necessary and sufficient condition for the convergence of the maximum likelihood estimator, as first conjectured by Carle and Strub (1978). Furthermore, when methods based on the likelihood function succeed in converging, the maximum likelihood estimate (MLE) will favor a small N_0 (Schnute, 1983). Secondly, asymptotic normality of the MLE fails in many situations, since the normal approximation holds only when N_0 and the total number of captures are large (Hirst, 1994).

1. Submitted article

To avoid the convergence problems of MLEs, it is possible to use Bayesian methods as an alternative for estimating both N_0 and τ (Bolfarine et al., 1992; Schwarz and Seber, 1999). One advantage of the Bayesian approach is that one can take into account prior knowledge of the parameters, when available. In the absence of prior information, the most common priors for τ are the uniform prior Beta(1,1), the Haldane prior Beta$(0, 0)$, or the Jeffreys prior for τ Beta$(\frac{1}{2}, \frac{1}{2})$. For N_0, the most natural choice is a flat prior (Bernardo, 1979a) which is improper and may be approximated by a uniform distribution of $\{0, 1, ..., n\}$, with large n.

The aim of this paper is to study the influence of prior distribution on the Bayesian inference of the population size N_0 for data collected by removal sampling. In Section 4.2.2, we establish some results on the limit behaviour of the likelihood and the profile likelihood of the removal sampling model. In Section 4.2.3, we consider a Bayesian inference for τ and N_0. First, we give necessary and sufficient conditions for the hyperparameters of the prior distributions in order to have proper posteriors distributions and well-defined estimates for N_0. Then, we consider proper vague priors and study the limiting behaviour of the posterior estimates based on the mean and the median of the marginal posterior distribution. In Section 4.2.4, we illustrate the theoretical results with simulations and case studies. In Section 4.2.5 we discuss the choice of the priors and give some recommendations.

4.2.2 Removal sampling likelihood and limit behaviour

4.2.2.1 Removal sampling likelihood

We consider k successive samplings in a closed population N_0 *i.e.* with no immigration, emigration, birth or death during the successive samplings, at a given point of observation. The aim of the experiment is to estimate the population size N_0 and incidentally the sampling rate τ.

Let X_i be the number of individuals captured at the i^{th} sampling. We assume that the probability of capture, τ, is constant across individuals and successive samplings and that individuals are captured independently. We assume that X_i follows a binomial distribution with population size $N_0 - \sum_{l=1}^{i-1} X_l$ and prob-

ability of capture τ. After k successive samplings, the vector of observations is $x = (x_1, x_2, ..., x_k)$, and the likelihood is

$$\begin{aligned} L((N_0, \tau); x) &= \prod_{i=1}^{k} P(X_i | x_1, ..., x_{i-1}) \\ &= \frac{N_0!}{(N_0 - T)! \prod_{i=1}^{k} x_i!} \tau^T (1-\tau)^{k(N_0-T)+d_0} \end{aligned} \tag{4.1}$$

where $T = \sum_{i=1}^{k} x_i$ and $d_0 = \sum_{i=1}^{k} (i-1)x_i$.

Note that only the part $(1-\tau)^{d_0}$ depends on the rank of sampling. When τ is close to 0, $(1-\tau)^{d_0}$ is close to 1, *i.e.* negligible.

4.2.2.2 Limit behavior of the likelihood function

Let us consider the limiting behaviour of the likelihood function when N_0 is large and τ is close to 0. Intuitively, when N_0 is large and τ is close to 0, the number of animals captured at each sampling is very low in comparison to N_0. So, the remaining population across successive samplings is approximatively constant and $X = (X_1, X_2, ..., X_k)$ behaves similarly to k independent random variables with a Binomial distribution $\mathrm{Bin}(\mathrm{N}_0, \tau)$. Moreover, the $\mathrm{Bin}(\mathrm{N}_0, \tau)$ distribution can be approximated by a Poisson distribution with parameter $N_0\tau$. Therefore the likelihood of the removal sampling model can be approximated by the likelihood of k independent Poisson distributions with parameter $N_0\tau$.

Proposition 4.2.1. *Assume that τ goes to 0, N_0 goes to $+\infty$ and there exists $\lambda_0 > 0$ such that $N_0\ \tau$ goes to λ_0. Then,*

$$\lim_{\substack{N_0 \to +\infty \\ N_0 \tau \to \lambda_0}} L((N_0, \tau); x) = L_p(\lambda_0; x)$$

where $L_p(\lambda_0; x) = \prod_{i=1}^{k} e^{-\lambda_0} \frac{\lambda_0^{x_i}}{x_i!}$ is the likelihood of k independent Poisson distributions with parameter λ.

Proof. We have $L((N_0, \tau); x) = \frac{N_0!\, \tau^T (1-\tau)^{k(N_0-T)+d_0}}{(N_0-T)! \prod_{i=1}^{k} x_i!}$. We make the change of vari-

able $\lambda = N_0\tau$, and we study what happen when N_0 goes to $+\infty$. We have

$$L\left((N_0, \frac{\lambda}{N_0}); x\right) = \frac{N_0\,!}{(N_0 - T)!\, N_0^T} \times \left(1 - \frac{\lambda}{N_0}\right)^{-kT+d_0} \tag{4.2}$$

$$\times \frac{\lambda^T}{\prod_{i=1}^{k} x_i!} \times \left(1 - \frac{\lambda}{N_0}\right)^{kN_0} \tag{4.3}$$

where the two left-hand side terms of the line (2) go to 1 when N_0 goes to $+\infty$. And in the line (3), $\lim_{N_0\to+\infty} \left(1 - \frac{\lambda}{N_0}\right)^{kN_0} = e^{-k\lambda_0}$. So, $\lim_{N_0\to+\infty} L\left((N_0, \frac{\lambda}{N_0}); x\right) = \frac{\lambda^T}{\prod x_i!}\, e^{-k\lambda_0} = \prod_{i=1}^{k} e^{-\lambda_0}\, \frac{\lambda_0^{x_i}}{x_i!} = L_p(\lambda_0; x)$. □

We may note that the Poisson iid limit model is not identifiable with respect to N_0 and τ. Indeed, for a given value of $\lambda = N_0\tau$ there exists an infinite number of combinations of N_0 and τ.

4.2.2.3 Limit behavior of the profile likelihood

Consider now the profile likelihood $L((N_0, \widehat{\tau}(N_0)); x)$ where $\widehat{\tau}(N_0) = T\{k(N_0 - T) + T + d_0\}^{-1}$ maximizes the likelihood $L((N_0, \tau); x)$ for a given N_0. The maximum likelihood estimator $\widehat{N}_{0,ML}$ can be obtained by maximizing the profile likelihood. The maximum likelihood estimator of τ is therefore $\widehat{\tau} = \widehat{\tau}(\widehat{N}_0)$.

Similarly to Proposition 4.2.1, the following proposition shows that the profile likelihood converges at the maximum likelihood of a Poisson iid model.

Proposition 4.2.2. *The profile likelihood of the removal sampling model satisfies the following convergence:*

$$\lim_{N_0\to+\infty} L((N_0, \widehat{\tau}(N_0)); x) = L_p(\widehat{\lambda}_{ML}; x)$$

where $\widehat{\lambda}_{ML} = \frac{T}{k}$ *is the ML estimator for the Poisson iid model.*

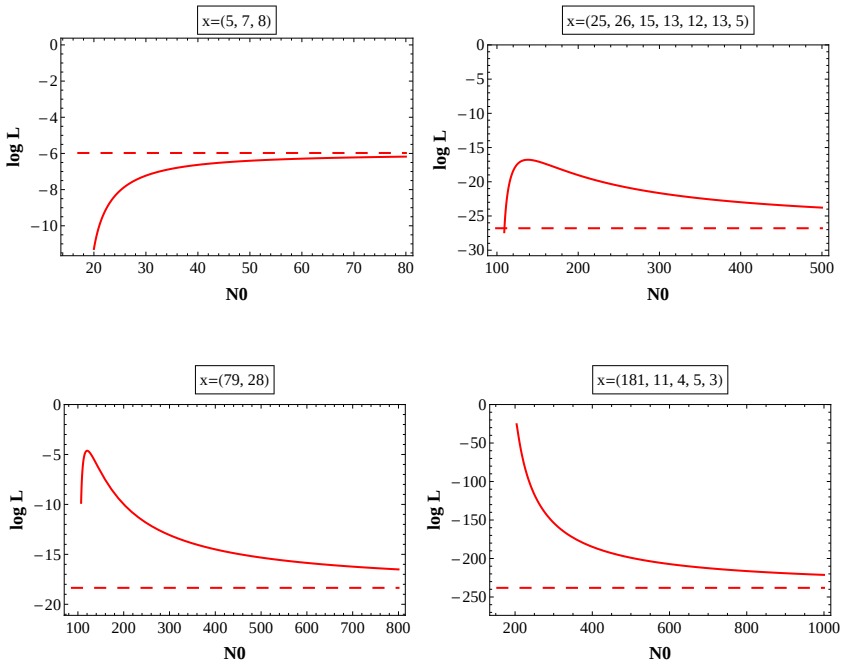

Figure 4.1: Profile log-likelihoods (continuous line) and their limit profile log-likelihoods (dashed lines) for the four data sets of Bedrick (1994).

Proof. We have

$$L((N_0, \widehat{\tau}(N_0)); x) = \frac{N_0!\, k^T}{(N_0 - T)!\, (k(N_0 - T) + T + d_0)^T} \tag{4.4}$$

$$\times \left(1 - \frac{T}{k(N_0 - T) + T + d_0}\right)^{-kT + d_0} \tag{4.5}$$

$$\times \frac{1}{\prod x_i!} \left(\frac{T}{k}\right)^T \times \left(1 - \frac{T}{k(N_0 - T) + T + d_0}\right)^{kN_0} \tag{4.6}$$

where the right-hand side of line (4) and line (5) go to 1 when N_0 goes to $+\infty$. And, in line (6), $\lim_{N_0 \to \infty} \left(1 - \frac{T}{k(N_0 - T) + T + d_0}\right)^{kN_0} = e^{-T}$. So $\lim_{N_0 \to \infty} L((N_0, \widehat{\tau}(N_0)); x)$ $= \frac{1}{\prod x_i!} \left(\frac{T}{k}\right)^T e^{-k\frac{T}{k}} = L_p(\widehat{\lambda}_{ML}; x)$ with $\widehat{\lambda}_{ML} = \frac{T}{k}$. □

Figure 4.1 displays four typical shapes of profile log-likelihood with their asymptotic limits. The first graphic give an example of non-convergent maximum likelihood estimator.

Proposition 4.2.1 show that the likelihood do not converge to 0 on the boundary of the parameter space, unlike most statistical models. In the next section, it will be seen that this feature have a significant impact in a bayesian context.

4.2.3 Bayesian analysis of removal sampling

We consider the following family of priors for (N_0, τ) :

$$\pi(N_0, \tau) \propto \frac{1}{N_0^c} \times \tau^{a-1}(1-\tau)^{b-1}. \tag{4.7}$$

Depending on the choice of the hyperparameters a, b and c, the behavior of the likelihood on the boundary established in Section 4.2.2 may lead to improper posterior distributions or divergent Bayes estimators. In order to get a good estimation of the abundance, we show that the prior distribution has to penalize small values of τ and/or large values of N_0. This is the case when $c > 0$ or $a > 1$. The posterior distribution is

$$\pi(N_0, \tau|x) \propto L((N_0, \tau); x)\ \pi(N_0, \tau). \tag{4.8}$$

4.2.3.1 Posterior analysis for N_0

We give here a necessary and sufficient condition for the hyperparameters a, b and c in order to ensure a proper posterior distribution, a well-defined Bayes estimator of N_0, and a well-defined posterior Bayes quadratic risk. First, we give a technical lemma.

Lemma 4.2.3. *For $a > 0$ and $b > 0$, we have*

$$\lim_{N_0 \to +\infty} N_0^a \int_0^1 \tau^{a-1}(1-\tau)^{b-1} L((N_0, \tau); x) \mathrm{d}\tau \ = \ K_{a,T} > 0$$

with $K_{a,T} = \int_0^{+\infty} \lambda^{a-1} L_p(\lambda; x) \mathrm{d}\lambda = \frac{1}{\prod_{i=1}^{k} x_i!} \frac{\Gamma(T+a)}{k^{T+a}}$.

Proof. Put $\lambda = N_0\tau$, we have $N_0^a \int_0^1 \tau^{a-1} (1-\tau)^{b-1} L((N_0, \tau); x) \mathrm{d}\tau = N_0^a \int_0^{+\infty} \left(\frac{\lambda}{N_0}\right)^{a-1} \left(1 - \frac{\lambda}{N_0}\right)^{b-1} L((N_0, \frac{\lambda}{N_0}); x) \mathbb{1}_{[0,N_0]}(\lambda) \frac{1}{N_0} \mathrm{d}\lambda = \int_0^{+\infty} \lambda^{a-1} \left(1 - \frac{\lambda}{N_0}\right)^{b-1}$

$L\left((N_0, \frac{\lambda}{N_0}); x\right)\ \mathbb{1}_{[0,N_0]}(\lambda)\ \mathrm{d}\lambda$. From Formula (4.1), for any λ, $\lambda^{a-1}\ \left(1-\frac{\lambda}{N_0}\right)^{b-1}$ $L\left((N_0, \frac{\lambda}{N_0}); x\right)\ \mathbb{1}_{[0,N_0]}(\lambda) \leqslant \lambda^{T+a-1}e^{-k\lambda}$ which is an integrable function. Moreover, from Proposition 4.2.1, $L\left((N_0, \frac{\lambda}{N_0}); x\right)\ \mathbb{1}_{[0,N_0]}(\lambda)$ converges to $L_p(\lambda; x)$ when N_0 goes to $+\infty$. So, from dominated convergence theorem, $\lim\limits_{N_0\to+\infty} \int_0^{+\infty} \lambda^{a-1}$ $\left(1-\frac{\lambda}{N_0}\right)^{b-1}\ L\left((N_0, \frac{\lambda}{N_0}); x\right)\ \mathbb{1}_{[0,N_0]}(\lambda)\ \mathrm{d}\lambda = \int_0^{+\infty}\ \lambda^{a-1}\ L_p(\lambda; x)\ \mathrm{d}\lambda = \frac{1}{\prod\limits_{i=1}^{k} x_i!} \int_0^{+\infty}$ $\lambda^{T+a-1}\ e^{-k\lambda}\ \mathrm{d}\lambda = \frac{1}{\prod\limits_{i=1}^{k} x_i!}\ \Gamma(T+a)k^{-(T+a)}$. □

Theorem 4.2.4. *Consider a prior π on (N_0, τ) whose density satisfies $\pi(N_0, \tau)$ $\propto \frac{1}{N_0^c} \times \tau^{a-1}(1-\tau)^{b-1}$. Then,*

1. *the posterior distribution $\pi(N_0, \tau|x)$ is proper if and only if $a + c > 1$,*
2. *the Bayes estimator of N_0, $\mathbb{E}_\pi(N_0|x)$, is finite if and only if $a + c > 2$,*
3. *the posterior Bayes quadratic risk for N_0 is finite if and only if $a + c > 3$.*

We may note that the conditions found here are similar to that found by Kahn (1987) for the Binomial model.

Proof.

1. We have $\sum\limits_{N_0>0} \pi(N_0|x) \propto \sum\limits_{N_0>0} \frac{1}{N_0^c} \int_0^1 L((N_0, \tau); x)\ \tau^{a-1}\ (1-\tau)^{b-1}\ \mathrm{d}\tau \propto \sum\limits_{N_0>0}$ $\frac{1}{N_0^{a+c}}\ N_0^a \int_0^1 L((N_0, \tau); x)\ \tau^{a-1}\ (1-\tau)^{b-1}\ \mathrm{d}\tau$. From Lemma 4.2.3, $\lim\limits_{N_0\to+\infty} N_0^a$ $\int_0^1 L((N_0, \tau); x)\ \tau^{a-1}\ (1-\tau)^{b-1}\ \mathrm{d}\tau = K_{a,T}$. So, $\sum\limits_{N_0>0} \pi(N_0|x)$ converges if $a + c > 1$.
2. Similarly to the point 1, $\mathbb{E}_\pi(N_0|x) \propto \sum\limits_{N_0>0} \frac{1}{N_0^{a+c-1}}\ N_0^a \int_0^1 L((N_0, \tau); x)\ \tau^{a-1}$ $(1-\tau)^{b-1}\ \mathrm{d}\tau$. Thus $\mathbb{E}_\pi(N_0|x) < +\infty$ if and only of $a + c - 1 > 1$, *i.e.* $a + c > 2$.
3. We have $\mathbb{E}_\pi\left((N_0 - \mathbb{E}(N_0|x))^2|x\right) = \mathbb{E}_\pi(N_0^2|x) - \mathbb{E}_\pi(N_0|x)^2$. But $\mathbb{E}_\pi(N_0^2|\ x)$ $\propto \sum\limits_{N_0>0} \frac{1}{N_0^{a+c-2}}\ N_0^a \int_0^1 L((N_0, \tau); x)\ \tau^{a-1}\ (1-\tau)^{b-1}\ \mathrm{d}\tau$. So, similarly to point 1 and 2, $\mathbb{E}_\pi\ \left((N_0 - \mathbb{E}(N_0|x))^2|x\right) < +\infty$ if and only if $a + c > 3$.

□

We notice that the conditions in Theorem 4.2.4 do not depend on b. Indeed, the improperness of the posterior distribution or the divergence of the estimator

comes from the behavior of the likelihood function when N_0 tends to $+\infty$ and τ tends to 0. This corresponds to $(1-\tau)^{b-1}$ tending to 1, whatever the value of $b > 0$.

4.2.3.2 Limiting behavior of sequences of proper priors

When no prior information is available, it is common to use a flat prior for both N_0 and τ which corresponds to the prior (4.7) with $c = 0$ and $a = b = 1$. However, we saw in section 4.2.3.1 that this prior leads to an improper posterior distribution. The usual way to obtain proper posterior distributions is to approximate the flat prior $\pi(N_0) \propto 1$ on N_0 by $\pi_n(N_0) \propto \mathbb{1}_{\{1 \leqslant N_0 \leqslant n\}}$ for large n. More generally, we can use any sequence Π_n of proper prior distributions that approximates the flat prior. The aim of this section is to show that, if the limit of the sequence of posterior distributions is improper, then, as expected, the sequence of Bayesian estimators diverges.

In the following, we use the definition of approximation proposed by Bioche and Druilhet (2015): a sequence $\{\Pi_n\}_n$ of proper priors is said to approximate an improper prior Π if there exists positive real numbers $\{a_n\}_n$ such that $\lim_{n\to+\infty} a_n \Pi_n(\phi) = \Pi(\phi)$ for any continuous function with compact support ϕ or in the discrete case, such that $\lim_{n\to+\infty} a_n \Pi_n(N_0) = \Pi(N_0)$. It can be shown that the limit is unique within a scalar factor and that the posterior distribution sequences $\Pi_n(N_0|x)$ also converge to $\Pi(N_0|x)$.

Lemma 4.2.5. *Let $\Pi_n(N_0, \tau) = \Pi_n^{(1)}(N_0) \times \Pi^{(2)}(\tau)$ where $\{\Pi_n^{(1)}\}_n$ is a sequence of proper priors on N_0 which approximates an improper prior $\Pi^{(1)}$ and $\Pi^{(2)}$ is a proper prior on τ. Define $\Pi(N_0|x) = \int_0^1 L((N_0, \tau); x)\Pi(N_0)\pi^{(2)}(\tau)\mathrm{d}\tau$. Then, the sequence $\{\Pi_n(N_0|x)\}$ of marginal posterior distributions on N_0 approximates $\Pi(N_0|x)$.*

Proof. Since $\{\Pi_n^{(1)}\}_n$ approximates $\Pi^{(1)}$, there exists $\{a_n\}_n$ such that for all N_0, $\lim_{n\to+\infty} a_n \Pi_n^{(1)}(N_0) = \Pi^{(1)}(N_0)$. Put $b(N_o) = \int L((N_0, \tau); x)\pi^{(2)}(\tau)\mathrm{d}\tau$ and $b_n = \sum_{N_0} \Pi_n(N_o)\, b(N_o)$. We have $\Pi_n(N_0|x) = b_n^{-1}\Pi_n(N_o)\, b(N_o)$. Therefore, $\lim_{n\to+\infty} a_n\, b_n \Pi_n(N_0|x) = b(N_0)\, \Pi(N_0) \propto \Pi(N_0|x)$. □

Now we state that if the marginal posterior $\Pi(N_0|x)$ is improper, then the sequences of posterior means and medians of $\Pi_n(N_0|x)$ diverge.

Proposition 4.2.6. *Assume that:*

1. $\Pi_n(N_0, \tau) = \Pi_n^{(1)}(N_0) \times \Pi^{(2)}(\tau)$ *where* $\Pi^{(2)}$ *is a proper prior and* $\{\Pi_n^{(1)}\}_n$ *is a sequence of proper priors on* N_0 *which approximates an improper prior* $\Pi^{(1)}$,
2. *the limit posterior distribution on* N_0, $\Pi(N_0|x)$, *is improper.*

Then,

a. $\lim\limits_{n\to+\infty} \mathbb{E}_{\Pi_n}(N_0|x) = +\infty$,

b. $\lim\limits_{n\to+\infty} \text{med}_{\Pi_n}(N_0|x) = +\infty$.

Proof.

a. From Lemma 4.2.5 and from asumption 1, $\{\Pi_n(N_0|x)\}_n$ is a sequence of probabilities which approximates $\Pi(N_0|x)$. From Proposition 2.6 by Bioche and Druilhet (2015), when a sequence of probabilities is used to approximate an improper prior, the mass tends to concentrate outside any compact set. For a discrete parameter, a compact set is a finite set. For all $A > 0$, we denote by C_A the set $\{0, \cdots, A\}$, and we have $\lim\limits_{n\to+\infty} \Pi_n(C_A^c|x) = 1$. So, for any $A > 0$, there exists n_A^* such that for $n > n_A^*$, $\Pi_n(C_A^c|x) \geqslant \frac{1}{2}$. We have $\mathbb{E}_{\Pi_n}(N_0|x) \geqslant \sum\limits_{N_0>A} N_0\Pi_n(N_0|x) \geqslant A\Pi_n(C_A^c|x)$. So, for $n > n_A^*$, $\mathbb{E}_{\Pi_n}(N_0|x) \geqslant \frac{A}{2}$. The result follows.

b. Similarly to the proof of the point a., we denote by C_A the set $\{0, \ldots, A\}$ and we can state that for any $A > 0$, there exists n_A^* such that for $n > n_A^*$, $\Pi_n(C_A^c|x) > \frac{1}{2}$. Then, for $n > n_A^*$, $\text{med}_{\Pi_n}(N_0|x) \geqslant A$. The result follows.

□

4.2.4 Case and simulation studies

The theoretical approach showed that to have good estimates of N_0, the prior distribution must penalize large values of N_0 and/or small values of τ, which corresponds to a large value of $a + c$ for the prior (4.7). Theorem 4.2.4 states that necessarily $a + c > 2$. This means in particular that we cannot simultaneously use non-informative priors for N_0 and τ. Here, we study the behavior of Bayesian estimates of N_0 according to the values of hyperparameters a and c through simulations. We also consider real data sets.

4.2.4.1 Simulation studies

We consider several scenarios, in which $N_0 = 50$ or 500 and $\tau = 0.1$, 0.3 or 0.5. For each scenario, we consider several values for a, b and c for the prior. The resulting estimators are compared using with the relative root mean square error (RRMSE) frequentist criterion which allows comparisons between senarios. The RRMSE of an estimator $\hat{N}_0$ of N_0 is defined by

$$\text{RRMSE}(\hat{N}_0) = \frac{\sqrt{\mathbb{E}((\hat{N}_0 - N_0)^2 | N_0, \tau)}}{N_0}$$

and similarly for τ. Following Pollock et al. (1990), a rough rule of thumb is that a study that provides a RRMSE smaller than 0.2 is reasonable.

Table 4.1: Ratio of the root of Mean Square Error (RRMSE) values of estimates if the population size N_0 and the sampling rate τ estimates according to the choice of prior distributions for N_0 and τ. Median and mean correspond, respectively, to the estimator based on the median and the mean of the posterior distribution.

τ	Prior		$a+c$	$N_0 = 50$				$N_0 = 500$			
				$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$		$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$	
	N_0	τ		$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$
0.1	flat	$\beta(3,3)$	3	0.423	1.889	0.538	1.853	0.385	0.692	0.321	0.683
	$1/N_0$	$\beta(2,2)$	3	0.451	2.034	0.562	1.995	0.363	0.727	0.330	0.720
	$1/N_0^2$	$\beta(1,1)$	3	0.469	2.106	0.577	2.070	0.360	0.734	0.334	0.727
	flat	$\beta(4,4)$	4	0.510	2.208	0.581	2.187	0.305	0.867	0.374	0.856
	$1/N_0$	$\beta(3,3)$	4	0.536	2.344	0.600	2.319	0.318	0.899	0.386	0.889
	$1/N_0^2$	$\beta(2,2)$	4	0.553	2.413	0.612	2.391	0.322	0.907	0.389	0.897
	$1/N_0^3$	$\beta(1,1)$	4	0.566	2.472	0.623	2.449	0.325	0.913	0.393	0.903
	flat	$\beta(5,5)$	5	0.557	2.435	0.604	2.422	0.366	1.019	0.420	1.007
	$1/N_0$	$\beta(4,4)$	5	0.577	2.558	0.620	2.542	0.377	1.047	0.429	1.037
	$1/N_0^2$	$\beta(3,3)$	5	0.590	2.621	0.630	2.605	0.380	1.054	0.433	1.044
	$1/N_0^3$	$\beta(2,2)$	5	0.602	2.680	0.639	2.665	0.384	1.061	0.436	1.051
	$1/N_0^2$	$\beta(4,4)$	6	0.612	2.777	0.642	2.765	0.425	1.179	0.464	1.170

Continued on next page

Table 4.1 – continued from previous page

τ	Prior		$a+c$	$N_0 = 50$				$N_0 = 500$			
				$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$		$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$	
	N_0	τ		$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$
	$1/N_0^3$	$\beta(3,3)$	6	0.622	2.833	0.649	2.822	0.428	1.187	0.467	1.178
	$1/N_0^3$	$\beta(4,4)$	7	0.634	2.953	0.658	2.944	0.460	1.297	0.490	1.289
0.3	flat	$\beta(3,3)$	3	0.512	0.321	0.257	0.339	0.193	0.164	0.141	0.165
	$1/N_0$	$\beta(2,2)$	3	0.451	0.347	0.241	0.364	0.166	0.163	0.136	0.163
	$1/N_0^2$	$\beta(1,1)$	3	0.432	0.364	0.236	0.381	0.164	0.164	0.134	0.163
	flat	$\beta(4,4)$	4	0.255	0.318	0.194	0.330	0.162	0.156	0.127	0.158
	$1/N_0$	$\beta(3,3)$	4	0.243	0.347	0.198	0.359	0.144	0.156	0.123	0.156
	$1/N_0^2$	$\beta(2,2)$	4	0.235	0.366	0.199	0.378	0.142	0.156	0.122	0.156
	$1/N_0^3$	$\beta(1,1)$	4	0.229	0.384	0.205	0.397	0.141	0.156	0.121	0.157
	flat	$\beta(5,5)$	5	0.193	0.329	0.183	0.340	0.138	0.150	0.116	0.152
	$1/N_0$	$\beta(4,4)$	5	0.189	0.358	0.191	0.368	0.129	0.150	0.113	0.151
	$1/N_0^2$	$\beta(3,3)$	5	0.189	0.376	0.197	0.386	0.127	0.151	0.113	0.152
	$1/N_0^3$	$\beta(2,2)$	5	0.191	0.394	0.202	0.404	0.126	0.152	0.112	0.152
	$1/N_0^2$	$\beta(4,4)$	6	0.180	0.389	0.201	0.397	0.117	0.148	0.107	0.149
	$1/N_0^3$	$\beta(3,3)$	6	0.185	0.407	0.207	0.415	0.116	0.149	0.106	0.150
	$1/N_0^3$	$\beta(4,4)$	7	0.188	0.420	0.212	0.427	0.108	0.148	0.102	0.149
0.5	flat	$\beta(3,3)$	3	0.293	0.191	0.166	0.189	0.036	0.061	0.030	0.061
	$1/N_0$	$\beta(2,2)$	3	0.252	0.190	0.152	0.190	0.030	0.061	0.029	0.061
	$1/N_0^2$	$\beta(1,1)$	3	0.235	0.191	0.144	0.192	0.030	0.061	0.029	0.061
	flat	$\beta(4,4)$	4	0.190	0.169	0.134	0.169	0.033	0.060	0.029	0.060
	$1/N_0$	$\beta(3,3)$	4	0.177	0.172	0.128	0.173	0.030	0.060	0.029	0.060
	$1/N_0^2$	$\beta(2,2)$	4	0.168	0.174	0.122	0.175	0.030	0.061	0.029	0.061
	$1/N_0^3$	$\beta(1,1)$	4	0.159	0.177	0.118	0.180	0.030	0.061	0.029	0.061
	flat	$\beta(5,5)$	5	0.158	0.156	0.120	0.156	0.032	0.060	0.029	0.060
	$1/N_0$	$\beta(4,4)$	5	0.148	0.158	0.113	0.159	0.030	0.060	0.029	0.060
	$1/N_0^2$	$\beta(3,3)$	5	0.140	0.160	0.110	0.162	0.029	0.060	0.029	0.060
	$1/N_0^3$	$\beta(2,2)$	5	0.133	0.164	0.106	0.167	0.029	0.061	0.029	0.061
	$1/N_0^2$	$\beta(4,4)$	6	0.124	0.149	0.103	0.151	0.029	0.060	0.029	0.060

Continued on next page

Table 4.1 – continued from previous page

τ	Prior		$a+c$	$N_0 = 50$				$N_0 = 500$			
				$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$		$\text{RRMSE}_{\text{mean}}$		$\text{RRMSE}_{\text{median}}$	
	N_0	τ		$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$	$\hat{N}_0$	$\hat{\tau}$
	$1/N_0^3$	$\beta(3,3)$	6	0.118	0.153	0.100	0.156	0.029	0.060	0.028	0.060
	$1/N_0^3$	$\beta(4,4)$	7	0.109	0.143	0.094	0.146	0.029	0.060	0.028	0.060

Simulations results are presented in Table 4.1. We can see that, for small values of τ (here $\tau = 0.1$), N_0 and τ are poorly estimated. As expected, estimators of N_0 have smaller RRMSE values when N_0 is large. We also see that choosing $a+c = 5$ or 6 gives better estimates of N_0 than $a + c = 3$ or 4. Indeed, a large value for c penalizes large values of N_0, while large values for a penalizes small values for τ inducing a shrunken estimator of N_0.

4.2.4.2 Case studies

We consider again the four data sets cited in Bedrick (1994) in order to compare estimations using likelihood and Bayesian approaches. The first data set comes from three trappings of mottled sculpin (which provides an illustration of divergent estimates of N_0), the second represents seven trappings of whitefish, the third gives the results of two trappings of trout and the last originates from five trappings of mayflies. We compare the maximum likelihood estimator and Bayesian estimators for several priors.

Table 4.2 – Continued on next page

data sets	Bayesian approach						Likelihood approach	
	Prior		$\hat{N}_0$		$\hat{\tau}$		$\hat{N}_{0MV}$	$\hat{\tau}_{MV}$
	N_0	τ	median	mean	median	mean		
x=(5,7,8)	flat	$\beta(3,3)$	47.70	70.04	0.17	0.18	∞	0
		$\beta(4,4)$	39.56	48.73	0.21	0.22		
	$1/N_0$	$\beta(3,3)$	38.00	46.09	0.22	0.23		
		$\beta(4,4)$	34.00	39.07	0.25	0.26		

data sets	Bayesian approach						Likelihood approach	
	Prior		$\hat{N}_0$		$\hat{\tau}$		$\hat{N}_{0MV}$	$\hat{\tau}_{MV}$
	N_0	τ	median	mean	median	mean		
	$1/N_0^2$	$\beta(3,3)$	33.00	38.02	0.25	0.26		
		$\beta(4,4)$	31.00	34.47	0.28	0.28		
x=(25,26,13,12,13,5)	flat	$\beta(3,3)$	138.30	143.03	0.20	0.20	115	0.24
		$\beta(4,4)$	137.00	140.58	0.20	0.20		
	$1/N_0$	$\beta(3,3)$	137.00	140.67	0.20	0.20		
		$\beta(4,4)$	136.00	138.52	0.21	0.21		
	$1/N_0^2$	$\beta(3,3)$	136.00	138.97	0.21	0.21		
		$\beta(4,4)$	134.00	137.00	0.21	0.21		
x=(79,28)	flat	$\beta(3,3)$	126.20	130.85	0.61	0.60	120	0.66
		$\beta(4,4)$	127.00	129.97	0.61	0.60		
	$1/N_0$	$\beta(3,3)$	125.00	127.81	0.62	0.61		
		$\beta(4,4)$	126.00	128.42	0.61	0.61		
	$1/N_0^2$	$\beta(3,3)$	124.00	126.73	0.62	0.62		
		$\beta(4,4)$	125.00	127.32	0.62	0.61		
x=(181,11,4,5,3)	flat	$\beta(3,3)$	204.30	205.26	0.80	0.80	204	0.82
		$\beta(4,4)$	204.30	205.02	0.80	0.80		
	$1/N_0$	$\beta(3,3)$	204.00	204.06	0.81	0.81		
		$\beta(4,4)$	204.00	204.07	0.81	0.81		
	$1/N_0^2$	$\beta(3,3)$	204.00	204.06	0.81	0.81		
		$\beta(4,4)$	204.00	204.07	0.81	0.81		

Table 4.2: Estimates of N_0 and τ using likelihood and Bayesian approaches based on the data sets cited in Bedrick (1994). Median and mean correspond, respectively, to the estimators based on the median and the mean of the posterior distribution in the Bayesian approach; $\hat{N}_{0MV}$ and $\hat{\tau}_{MV}$ correspond, respectively, to the estimators of N_0 and τ obtained with the maximum likelihood approach.

The results presented in Table 4.2 show that the Bayesian approach is able to give an estimator of N_0 for the four data sets, even for the first in which the maximum likelihood estimator diverges. For the second and third data sets, the Bayesian estimator of N_0 is greater than the maximum likelihood estimator, which

is closer to the total number of captures. This result is consistent with that of Schnute (1983), which showed that the maximum likelihood estimate favors small values of N_0. In the fourth data set, the Bayesian estimator of N_0 equals the maximum likelihood estimator and the total number of captures.

4.2.5 Conclusion

To estimate the abundance N_0 of an animal population using an unknown sampling rate τ, the removal method is an useful sampling design. When the true sampling rate is small *e.g.* less than 10%, Bayesian or frequentist estimation methods do not lead to good estimates of abundance (see *e.g.* Otis et al. (1978) or White et al. (1982)), except when accurate knowledge is available for τ. For larger sampling rates, the theoretical results and the simulation studies show that Bayesian methods lead to good estimates of abundance only when the prior distribution penalizes large values of N_0 and/or small values of τ. This means in particular that we cannot simultaneously use non-informative priors for N_o and τ. In practice, as we often lack precise knowledge on N_0 and τ we may use the prior (4.7). Theorem 4.2.4 states that necessarily, $a+c > 2$, but simulation studies show that $a + c > 4$ is preferable. We can also observe that the overall value of $a + c$ is more important than the specific allocation between a and c in order to have a good estimate of abundance.

Chapitre 5

From convergence on priors to logarithmic and expected logarithmic convergence of posteriors [1]

5.1 Introduction and notations

Wallace (1959); Stone (1965, 1970); Heath and Sudderth (1989) justify the use of improper priors by showing that the formal posteriors are suitable limit of posteriors obtained from proper priors. They all consider different convergence modes on the posterior distributions (see section 2.2). Berger et al. (2009) consider the logarithmic convergence and the expected logarithmic convergence of posteriors.

The aim of this article is to establish links between convergence of priors and logarithmic convergence of posteriors. We define a new convergence mode, the q-monotone convergence, which is a little more restrictive than the q-vague convergence. We show that the q-monotone convergence of priors implies the logarithmic convergence of posteriors. We also give some other sufficient conditions on priors to obtain the logarithmic convergence of posteriors. Theorem 1 by Berger et al. (2009) states the logarithmic convergence of posteriors only for sequences of priors

1. Draft article

obtained by truncation, we generalize this result to other approximating sequences of priors. The final section examines the expected logarithmic convergence of posteriors for observations from the location model. We also propose a generalization of a result of Berger et al. (2009).

Let X be a random variable and assume that $X|\theta \sim P_\theta$, $\theta \in \Theta$. We assume that Θ is a locally compact Hausdorff space that is second countable. This ensures that there exists a sequence of compact sets $\{\Theta_n\}_n$ such that $\Theta = \bigcup_n \Theta_n$ and $\Theta_n \subset \overset{\circ}{\Theta}_{n+1}$ where $\overset{\circ}{\Theta}_n$ is the interior of Θ_n. In practice, Θ is often in $\mathbb{R}$, $\mathbb{R}^p$, $p > 1$, or a countable set. It is assumed that probability distributions may be described through probability density functions, either in respect to Lebesgue measure or counting measure. We denote by π the density function of a measure Π. No distinction is made between a random quantity and the particular values that it may take. The conditional probability density of data $x \in \mathcal{X}$ given the parameter θ will be represented by $f(x|\theta)$ with $f(x|\theta) \geqslant 0$ and $\int_{\mathcal{X}} f(x|\theta)\mathrm{d}x = 1$. The posterior distribution of $\theta \in \Theta$ given x will be represented by $\pi(\theta|x)$, with $\pi(\theta|x) \geqslant 0$ and $\int_\Theta \pi(\theta|x)\mathrm{d}\theta = 1$. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums.

The logarithmic convergence or convergence in relative entropy is based on the Kullback-Leibler divergence (also called relative entropy).

Definition 5.1.1. *The Kullbabck-Leibler divergence between two probability densities π and $\widetilde{\pi}$ is defined by*

$$D(\widetilde{\pi}\|\pi) = \int_\Theta \widetilde{\pi}(\theta) \log\left(\frac{\widetilde{\pi}(\theta)}{\pi(\theta)}\right) \mathrm{d}\theta$$

provided the integral (or the sum) is finite.

The properties of $D(\widetilde{\pi}\|\pi)$ have been extensively studied (Gibbs, 1902; Shannon, 1948; Good, 1950, 1969; Kullback and Leibler, 1951; Chernoff, 1956; Jaynes, 1957, 1968; Kullback, 1959; Csiszár, 1967, 1975). We recall that for π and $\widetilde{\pi}$ probability densities, $D(\widetilde{\pi}\|\pi) \geqslant 0$.

Definition 5.1.2. *A sequence of probability density functions $\{\pi_n\}_n$ is said to*

converge logarithmically to a probability density function π if and only if

$$\lim_{n\to\infty} D(\pi_n\|\pi) = 0.$$

Berger et al. (2009) consider sequences of posteriors corresponding to priors obtained by truncation, that is:

Definition 5.1.3 (Berger et al. (2009))**.** *Consider a parametric model $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ and a strictly positive continuous function $\pi(\theta)$, such that $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$ for all $x \in \mathcal{X}$. An approximating compact sequence of parameter spaces is an increasing sequence of compact subsets of Θ, $\{\Theta_n\}_n$, converging to Θ. The corresponding sequence of posteriors with support on Θ_n, defined as $\{\pi_n(\theta|x)\}_n$, with*

$$\pi_n(\theta) = \frac{\pi(\theta)\mathbb{1}_{\Theta_n}(\theta)}{\int_{\Theta_n}\pi(\theta)\mathrm{d}\theta}$$

is called the approximating sequence of posteriors to the formal posterior $\pi(\theta|x)$.

They show that any approximating sequence of posteriors converges logarithmically to the formal posterior $\pi(\theta|x)$.

Theorem 5.1.4 (Berger et al. (2009), Theorem 1)**.** *Consider a model $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ and a strictly positive continuous function $\pi(\theta)$, such that $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$ for all $x \in \mathcal{X}$. For any approximating compact sequence of parameter spaces, the corresponding approximating sequence of posteriors converges logarithmically to the formal formal posterior $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.*

5.2 Generalization to other approximating sequences of priors

In this section, we will generalize Theorem 5.1.4 by showing the logarithmic convergence of posteriors for other approximating sequences of improper priors.

We introduce the concept of q-monotone convergence which is a derivative version of the q-vague convergence. This concept extends the notion of approximat-

ing sequences obtained by truncation to other increasing approximating sequences. First, we recall the definition of the q-vague convergence presented in Chapter 3:

Definition 5.2.1. *A sequence of positive Radon measures* $\{\Pi_n\}_n$ *is said to converge q-vaguely to a positive Radon measure* Π *if there exists a sequence of positive real numbers* $\{a_n\}_n$ *such that* $\{a_n\Pi_n\}_n$ *converges vaguely to* Π*.*

We recall that a sequence of prior measures cannot converge q-vaguely to more that one limit up to within a scalar factor. The q-monotone convergence is defined by:

Definition 5.2.2. *A sequence of positive Radon measures* $\{\Pi_n\}_n$ *is said to converge q-monotonically to the positive Radon measure* Π *if there exists a sequence of positive scalars* $\{a_n\}_n$ *such that* $\{a_n\pi_n\}_n$ *is a non-decreasing sequence which converges pointwise to* π*.*

As in the case of the q-vague convergence, we justify the use of the sequence $\{a_n\}_n$ in this definition by the fact that for $\alpha > 0$, Π and $\alpha\Pi$ give the same posterior distribution. We can note that Wallace (1959) also looked at sequences of priors up to within a scalar factor (see Proposition 2.2.1). It can also be shown that a sequence of prior measures cannot converge q-monotonically to more than one limit up to within a scalar factor.

The q-motonone convergence is stronger that the q-vague convergence.

Remark 5.2.3. *If a sequence of positive Radon measures* $\{\Pi_n\}_n$ *converges q-monotonically to a positive Radon measure* Π*, then* $\{\Pi_n\}_n$ *converges q-vaguely to* Π*.*

Any improper Radon measure can be approximated, in the sense of the q-monotone convergence, by a sequence of proper priors.

Remark 5.2.4. *For any Radon measure* Π*, and for any increasing sequence of compacts* $\{\Theta_n\}_n$ *which converges to* Θ*, the sequence of priors* $\{\Pi_n\}_n$ *defined by* $\pi_n(\theta) = c_n^{-1}\pi(\theta)\mathbb{1}_{\Theta_n}(\theta)$ *where* $c_n = \int_{\Theta_n}\pi(\theta)\mathrm{d}\theta$ *converges q-monotonically to* Π*. We just have to take* $a_n = c_n$ *in Definition 5.2.2.*

We give some examples of usual sequences of priors which converges q-monotonically to improper priors.

Example 5.2.5.

1. *The sequence of uniform distributions* $\{\mathcal{U}(\{0,\dots,n\})\}_n$ *converges q-monotonically to the counting measure.*
2. *The sequence of uniform distributions* $\{\mathcal{U}([-n,n])\}_n$ *converges q-monotonically to the Lebesgue measure.*
3. *The sequence of normal distributions* $\{\mathcal{N}(0,n)\}_n$ *converges q-monotonically to the Lebesgue measure.*
4. *The sequence of Beta distributions* $\{\text{Beta}(\frac{1}{n},\frac{1}{n})\}_n$ *on* $]0;1[$ *converges q-monotonically to the Haldane prior* $\Pi_H(\theta) = [\theta(1-\theta)]^{-1}$.

We now state that, with the same asumptions as Theorem 5.1.4 on the model and the prior density π; if a sequence of priors converges q-monotonically to an improper prior Π, the corresponding sequence of posteriors converges logarithmically to the formal posterior $\Pi(\theta|x)$.

Proposition 5.2.6. *Consider a model* $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ *and a strictly positive continuous function* $\pi(\theta)$*, such that* $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$ *for all* $x \in \mathcal{X}$*. Assume that there exists a sequence of probabilities* $\{\Pi_n\}_n$ *such that* $\{\Pi_n\}_n$ *converges monotonically to* Π*, then* $\{\pi_n(\theta|x)\}_n$ *converges logarithmically to* $\pi(\theta|x)$*.*

Proof.

$$
\begin{aligned}
D(\pi_n(.|x)\|\pi(.|x)) &= \int_\Theta \pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta \\
&= \int_\Theta \pi_n(\theta|x)\log\left(\frac{\pi_n(\theta)f(x|\theta)}{\int_\Theta f(x|\theta)\pi_n(\theta)\mathrm{d}\theta}\times\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\pi(\theta)f(x|\theta)}\right)\mathrm{d}\theta \\
&= \int_\Theta \pi_n(\theta|x)\log\left(\frac{a_n\pi_n(\theta)}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\times\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\pi(\theta)}\right)\mathrm{d}\theta \\
&= \int_\Theta \pi_n(\theta|x)\log\left(\frac{a_n\pi_n(\theta)}{\pi(\theta)}\times\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\right)\mathrm{d}\theta \\
&\leqslant \int_\Theta \pi_n(\theta|x)\log\left(\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\right)\mathrm{d}\theta
\end{aligned}
$$

since $0 \leqslant \frac{a_n\pi_n(\theta)}{\pi(\theta)} \leqslant 1$ for all θ and all n and $0 < \frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}$. From the monotone convergence theorem, $\lim_{n\to\infty} \int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$ so there exists N such that for $n > N$, $\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta} \leqslant 1+\varepsilon$. Then, for all $\varepsilon > 0$, there exists N such that for all $n > N$,

$$\begin{aligned} 0 \leqslant D(\pi_n(.|x)\|\pi(.|x)) &\leqslant \int_\Theta \pi_n(\theta|x)\log\left(\frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\right)\mathrm{d}\theta \\ &\leqslant \int_\Theta \pi_n(\theta|x)\log(1+\varepsilon)\mathrm{d}\theta \\ &\leqslant \log(1+\varepsilon) \end{aligned}$$

Consequently, $\{\pi_n(\theta|x)\}_n$ converges logarithmically to $\pi(\theta|x)$. □

The two last sequences considered in Example 5.2.5 do not satisfy hypothesis of Theorem 5.1.4 but satisfy these of Proposition 5.2.6 provided that the formal posterior is well-defined. So, we have proposed a generalization since Berger et al. (2009) were limited to sequences of priors obtained by truncation.

A bordeline case can be illustrate with the sequence considered in *4.* in Example 5.2.5. Consider the binomial model $X|\theta \sim \mathrm{Bin}(N,\theta)$ and the sequence of priors $\Pi_n = \mathrm{Beta}\left(\frac{1}{n},\frac{1}{n}\right)$. On $]0,1[$, the density of Π_n with respect to the Lebesgue measure $\lambda_\mathbb{R}$ is given by

$$\pi_n(\theta) = \frac{1}{B\left(\frac{1}{n},\frac{1}{n}\right)}\theta^{\frac{1}{n}-1}(1-\theta)^{\frac{1}{n}-1}\mathbb{1}_{]0,1[}(\theta) \tag{5.1}$$

where $B(x,y)$ is the Beta function, that is, $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$. As shown in Example 5.2.5, $\{\mathrm{Beta}\left(\frac{1}{n},\frac{1}{n}\right)\}_n$ converges q-monotonically to the Haldane prior Π_H defined by $\pi_H(\theta) = [\theta(1-\theta)]^{-1}$. For $0 < x < N$, it can be shown that $\{\pi_n(\theta|x)\}_n$ converges logarithmically to $\pi_H(\theta|x)$. However, for $x = 0$ and $x = N$, we have $\pi_H(\theta|x=0) = \theta^{-1}(1-\theta)^{N-1}$ and $\pi_H(\theta|x=N) = \theta^{N-1}(1-\theta)^{-1}$ which are improper measures and the logarithmic convergence is defined only for two probability measures. This refers to the asumption "$\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$ for all $x \in \mathcal{X}$". To continue the discussion on this example, it can be shown (see section 3.1.6.2) that the sequence $\{\mathrm{Beta}\left(\frac{1}{n},\frac{1}{n}\right)\}_n$ converges q-vaguely to $\Pi_{\{0,1\}} = \frac{1}{2}(\delta_0+\delta_1)$ on $[0,1]$. In this case, we consider the density given by Equation (5.1) with respect to the measure $\lambda_\mathbb{R} + \delta_0 + \delta_1$. However, the density of the limiting measure $\Pi_{\{0,1\}}$

with respect to the measure $\lambda_{\mathbb{R}} + \delta_0 + \delta_1$ is not a strictly positive function so we are not in the context of Proposition 5.2.6.

The q-monotone convergence of a sequence of priors is sufficient to provide the logarithmic convergence of a the corresponding sequence of posteriors but is not necessary. Proposition 5.2.7 gives some other sufficient conditions on a sequence of priors to entail the logarithmic convergence of the corresponding sequence of posteriors. This proposition shows that sequences of priors do not necessary need to be increasing sequence to entail the logarithmic convergence of posteriors.

Proposition 5.2.7. *Consider a parametric model $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$. Let Π be a positive Radon measure such that $\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta < +\infty$. Assume that there exist a sequence of probability measures $\{\Pi_n\}_n$ and a sequence of positive scalars $\{a_n\}_n$ such that:*

1. *$\{a_n\pi_n\}_n$ converges pointwise to π,*
2. *$\left\{\frac{a_n\pi_n}{\pi}\right\}_n$ converges to 1 uniformly on compact sets,*
3. *there exists a function $g : \Theta \longrightarrow \mathbb{R}^+$ such that $\theta \longmapsto f(x|\theta)g(\theta)$ is Lebesgue-integrable for all x and $a_n\pi_n(\theta) < g(\theta)$ for all $\theta \in \Theta$.*

Then, $\{\pi_n(\theta|x)\}_n$ converges logarithmically to $\pi(\theta|x)$.

Proof. Let $\{\Theta_l\}_l$ be a sequence of compact sets such that $\Theta_l \subset \overset{\circ}{\Theta}_{l+1}$ and $\bigcup_l \Theta_l = \Theta$. We have

$$\int_\Theta \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) \mathrm{d}\theta = \lim_{l\to\infty} \int_{\Theta_l} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) \mathrm{d}\theta.$$

For each l,

$$\int_{\Theta_l} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) \mathrm{d}\theta = \int_{\Theta_l} \pi_n(\theta|x) \log\left(\frac{a_n\pi_n(\theta)}{\pi(\theta)} \frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\right) \mathrm{d}\theta.$$

Let us study $\int_{\Theta_l} \pi_n(\theta|x) \log\left(\frac{a_n\pi_n(\theta)}{\pi(\theta)} \times \frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta}\right) \mathrm{d}\theta$.

– It is assumed that $\left\{\frac{a_n\pi_n}{\pi}\right\}_n$ converges to 1 uniformly on compact sets. So for $\varepsilon > 0$, for all l there exists $N_{1,l}$ such that for $n > N_{1,l}$, $\sup_{\theta\in\Theta_l} \left|\frac{a_n\pi_n(\theta)}{\pi(\theta)} - 1\right| \leqslant \varepsilon$.

– From assumption *1.* $\{a_n\pi_n(\cdot)\}_n$ converges pointwise to $\pi(\cdot)$, so $\{a_n\pi_n(\cdot) f(x|\cdot)\}_n$ converges pointwise to $\pi(\cdot)f(x|\cdot)$ for each x. From assumption *3.* and by dominated convergence theorem,

$$\lim_{n\to\infty}\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta.$$

So for $\varepsilon > 0$, there exists N_2 such that for $n > N_2$,

$$1-\varepsilon \leqslant \frac{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)a_n\pi_n(\theta)\mathrm{d}\theta} \leqslant 1+\varepsilon$$

Then, for $\varepsilon > 0$, there exists $N_l = \max(N_{1,l}, N_2)$ such that

$$\int_{\Theta_l}\pi_n(\theta|x)\log(1-\varepsilon)^2\mathrm{d}\theta \leqslant \int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta \leqslant \int_{\Theta_l}\pi_n(\theta|x)\log(1+\varepsilon)^2\mathrm{d}\theta$$

$$\log(1-\varepsilon)^2\int_{\Theta_l}\pi_n(\theta|x)\mathrm{d}\theta \leqslant \int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta \leqslant \log(1+\varepsilon)^2\int_{\Theta_l}\pi_n(\theta|x)\mathrm{d}\theta.$$

Since $\Pi_n(\cdot|x)$ is a probability measure, $0 \leqslant \int_{\Theta_l}\pi_n(\theta|x)\mathrm{d}\theta \leqslant 1$. So,

$$\log(1-\varepsilon)^2 \leqslant \log(1-\varepsilon)^2\int_{\Theta_l}\pi_n(\theta|x)\mathrm{d}\theta$$

and

$$\log(1+\varepsilon)^2\int_{\Theta_l}\pi_n(\theta|x)\mathrm{d}\theta \leqslant \log(1+\varepsilon)^2.$$

Thus, for $\varepsilon > 0$, for all l, there exists N_l such that for $n > N_l$

$$\log(1-\varepsilon)^2 \leqslant \int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta \leqslant \log(1+\varepsilon)^2.$$

Consequently, for all l,

$$\lim_{n\to\infty}\int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta = 0.$$

And we have

$$\begin{aligned}\lim_{n\to\infty}\int_\Theta \pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta &= \lim_{n\to\infty}\lim_{l\to\infty}\int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta\\ &= \lim_{l\to\infty}\lim_{n\to\infty}\int_{\Theta_l}\pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right)\mathrm{d}\theta\\ &= \lim_{l\to\infty} 0\\ &= 0.\end{aligned}$$

The result follows. □

Example 5.2.8.

- *Sequences defined by* $\Pi_n = \mathrm{Gamma}(\alpha_n, \beta_n)$ *with* $\lim_{n\to\infty}(\alpha_n, \beta_n) = (0, 0)$ *are used to approximate* $\Pi = \theta^{-1}\mathbb{1}_{\theta>0}\mathrm{d}\theta$. *In fact,* $\{\mathrm{Gamma}(\alpha_n, \beta_n)\}_n$ *converges q-vaguely to* $\theta^{-1}\mathbb{1}_{\theta>0}\mathrm{d}\theta$ *but we have seen that the q-vague convergence of priors is not sufficient to imply the logarithmic convergence of posteriors. However, the sequence* $\{\mathrm{Gamma}(\alpha_n, \beta_n)\}_n$ *satisfies assumptions of proposition 5.2.7. Indeed,*
  - *for* $a_n = \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}$, $a_n\pi_n(\theta) = \theta^{\alpha_n-1}e^{-\beta_n\theta}$ *which converges pointwise to* θ^{-1} *when* $\{\alpha_n\}_n$ *and* $\{\beta_n\}_n$ *tend to* 0.
  - $\frac{a_n\pi_n(\theta)}{\pi(\theta)} = \theta^{\alpha_n}e^{-\beta_n\theta}$ *converges to* 1 *uniformly on compact sets.*
  - $g(\theta) = \frac{1}{\theta}\mathbb{1}_{]0,1[}(\theta) + \mathbb{1}_{[1,+\infty[}(\theta)$ *satisfies assumption 3. of proposition 5.2.7 for a Poisson model.*

  Then, for a Poisson model, the sequence of posteriors $\{\pi_n(\theta|x)\}_n$ *converges logarithmically to* $\pi(\theta|x)$.
- *Sequences defined by* $\Pi_n = \mathrm{Gamma}(\alpha_n, 1)$ *with* $\lim_{n\to\infty}\alpha_n = 0$ *are used to approximate* $\Pi = \theta^{-1}e^{-\theta}\mathbb{1}_{\theta>0}\mathrm{d}\theta$. *We have*
  - *for* $a_n = \frac{1}{\Gamma(\alpha_n)}$, $a_n\pi_n(\theta) = \theta^{\alpha_n-1}e^{-\theta}$ *which converges pointwise to* $\theta^{-1}e^{-\theta}$ *when* $\{\alpha_n\}_n$ *tends to* 0.
  - $\frac{a_n\pi_n(\theta)}{\pi(\theta)} = \theta^{\alpha_n}$ *converges to* 1 *uniformly on compact sets.*
  - $g(\theta) = \frac{1}{\theta}\mathbb{1}_{]0,1[}(\theta) + \mathbb{1}_{[1,+\infty[}(\theta)$ *satisfies assumption 3. of proposition 5.2.7 for a Poisson model.*

  Then, for a Poisson model, the sequence of posteriors $\{\pi_n(\theta|x)\}_n$ *converges logarithmically to* $\pi(\theta|x)$.

5.3 Expected logarithmic convergence

The logarithmic convergence is a pointwise convergence. Berger et al. (2009) consider a stronger notion of convergence which guarantees that the approximating posteriors are accurate in a global sense over x.

Definition 5.3.1 (Berger et al. (2009))**.** *Consider a parametric model* $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ *and a strictly positive continuous function* $\pi(\theta)$*. The sequence of posterior probability densities* $\{\pi_n(\theta|x)\}_n$ *is said to converge expected logarithmically to a posterior probability density* $\pi(\theta|x)$ *if*

$$\lim_{n\to\infty} \int_{\mathcal{X}} D(\pi_n(.|x)\|\pi(.|x))p_n(x)\mathrm{d}x = 0$$

where $p_n(x) = \int_\Theta f(x|\theta)\pi_n(\theta)\mathrm{d}\theta$*.*

Berger et al. (2009) define a permissible prior for a model $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ as a strictly positive continuous function $\pi(\theta)$ such that:

1. for all $x \in \mathcal{X}$, $\pi(\theta|x)$ is proper;
2. for an increasing sequence of compact sets $\{\Theta_n\}_n$ such that $\bigcup \Theta_n = \Theta$, the corresponding posterior sequence (obtained by truncation on the priors) is expected logarithmically convergent to $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

Once more, they focus solely on sequences obtained by truncation. They show that, for one observation from a location model, the objective prior $\pi(\theta) = 1$ is permissible under mild conditions.

We revisit the definition of a permissible prior for a model $\mathcal{M} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ by proposing an alternative to condition 2:

2'. there exists a sequence of proper priors $\{\Pi_n\}_n$ such that the corresponding posterior sequence is expected logarithmically convergent to $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

In Proposition 5.3.2, we state that, for another type of approximant sequences of priors, for one observation from a location model, the objective prior $\pi(\theta) = 1$ is permissible, for the new definition, under some conditions.

Proposition 5.3.2. *Consider the model* $\mathcal{M} = \{f(x-\theta), \theta \in \mathbb{R}, x \in \mathbb{R}\}$*, where* $f(t)$ *is a density function on* $\mathbb{R}$ *integrable, continuous at* 0 *and such that* $f(0) > 0$*.*

Assume that the Fourier transform of f is of the form $\exp(\phi(\xi))$ and that there exists a function h such that $\phi(\alpha\xi) + \phi(n\xi) = \phi(h(\alpha, n)\xi)$ with $\lim_{n\to\infty} \frac{h(\alpha,n)}{n} = 1$ for $\alpha > 0$. Then, the sequence of priors $\{\Pi_n\}_n$ defined by $\pi_n(\theta) = \frac{1}{n} f(\frac{\theta}{n})$ provides a sequence of posteriors which is expected logarithmically convergent to the formal posterior corresponding to the improper prior $\pi(\theta) = 1$.

Proof. By the invariance of the model $p(x) = \int_{\mathbb{R}} f(x-\theta)\pi(\theta)\mathrm{d}\theta = 1$ and $\pi(\theta|x) = f(x-\theta)$. Then,

$$
\begin{aligned}
&\int_{\mathbb{R}}\int_{\mathbb{R}} \pi_n(\theta|x) \log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) p_n(x)\mathrm{d}\theta\mathrm{d}x \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{\pi_n(\theta)f(x-\theta)}{p_n(x)} \log\left(\frac{\pi_n(\theta)f(x-\theta)}{p_n(x)} \times \frac{1}{f(x-\theta)}\right) p_n(x)\mathrm{d}\theta\mathrm{d}x \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \pi_n(\theta)f(x-\theta) \log\left(\frac{\pi_n(\theta)}{p_n(x)}\right) \mathrm{d}\theta\mathrm{d}x \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \pi_n(\theta)f(x-\theta)\log(\pi_n(\theta))\mathrm{d}\theta\mathrm{d}x - \int_{\mathbb{R}}\int_{\mathbb{R}} \pi_n(\theta)f(x-\theta)\log(p_n(x))\mathrm{d}\theta\mathrm{d}x \\
&= \int_{\mathbb{R}} \pi_n(\theta)\log(\pi_n(\theta)) \int_{\mathbb{R}} f(x-\theta)\mathrm{d}x\mathrm{d}\theta - \int_{\mathbb{R}} \log(p_n(x)) \int_{\mathbb{R}} \pi_n(\theta)f(x-\theta)\mathrm{d}\theta\mathrm{d}x \\
&= \int_{\mathbb{R}} \pi_n(\theta)\log(\pi_n(\theta))\mathrm{d}\theta - \int_{\mathbb{R}} p_n(x)\log(p_n(x))\mathrm{d}x.
\end{aligned}
$$

By definition $p_n(x) = \int_{\mathbb{R}} f(x-\theta)\pi_n(\theta)\mathrm{d}\theta$. So, $p_n(x) = (f * \pi_n)(x)$ where $f * \pi_n$ denotes the convolution of f and π_n. If we denote by $\mathcal{F}$ the Fourier transform, we have

$$\mathcal{F}(p_n) = \mathcal{F}(f * \pi_n) = \mathcal{F}(f) \times \mathcal{F}(\pi_n).$$

From properties of the Fourier transform, $\hat{\pi}_n(\xi) = \hat{f}(n\xi)$. So, $\hat{p}_n(\xi) = \hat{f}(\xi)\hat{f}(n\xi)$. Since, $\hat{f}(\xi) = \exp(\phi(\xi))$,

$$\hat{p}_n(\xi) = \exp(\phi(\xi))\exp(\phi(n\xi)) = \exp(\phi(\xi) + \phi(n\xi)) = \exp(\phi(h(1,n)\xi)).$$

From the inverse Fourier transform $f(x) = \mathcal{F}^{-1}(\hat{f})(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{f}(\xi)\, e^{+\mathrm{i}\xi x}\,\mathrm{d}\xi$, we

obtain $p_n(x) = \frac{1}{h(1,n)} f\left(\frac{x}{h(1,n)}\right)$. Thus, we have

$$\begin{aligned}\int_{\mathbb{R}} \pi_n(\theta)\log(\pi_n(\theta))\mathrm{d}\theta - \int_{\mathbb{R}} p_n(x)\log(p_n(x))\mathrm{d}x \\ &= \int_{\mathbb{R}} \frac{1}{n} f\left(\frac{\theta}{n}\right) \log\left(\frac{1}{n} f\left(\frac{\theta}{n}\right)\right) \mathrm{d}\theta \\ &\quad - \int_{\mathbb{R}} \frac{1}{h(1,n)} f\left(\frac{x}{h(1,n)}\right) \log\left(\frac{1}{h(1,n)} f\left(\frac{x}{h(1,n)}\right)\right) \mathrm{d}x \\ &= \int_{\mathbb{R}} f(\eta)\log\left(\frac{1}{n} f(\eta)\right)\mathrm{d}\eta - \int_{\mathbb{R}} f(y)\log\left(\frac{1}{h(1,n)} f(y)\right)\mathrm{d}y\end{aligned}$$

by the changes of variable $\eta = \frac{\theta}{n}$ and $y = \frac{x}{h(1,n)}$.
So,

$$\begin{aligned}&\int_{\mathbb{R}} \pi_n(\theta)\log(\pi_n(\theta))\mathrm{d}\theta - \int_{\mathbb{R}} p_n(x)\log(p_n(x))\mathrm{d}x \\ &= \int_{\mathbb{R}} f(t)\log\left(\frac{h(1,n)}{n}\right)\mathrm{d}t = \log\left(\frac{h(1,n)}{n}\right).\end{aligned}$$

It's assumed that $\lim_{n\to\infty} \frac{h(\alpha,n)}{n} = 1$. Thus,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \int_{\mathbb{R}} \pi_n(\theta|x)\log\left(\frac{\pi_n(\theta|x)}{\pi(\theta|x)}\right) p_n(x)\mathrm{d}\theta\mathrm{d}x = 0.$$

□

We give two examples of classical location models which satisfy hypothesis of Proposition 5.3.2.

Example 5.3.3.

– *Consider the Gaussian model, then we have* $f(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$. *The corresponding sequence of priors given by Proposition 5.3.2 is* $\Pi_n = \mathcal{N}(0, n^2)$. *The Fourier transform of* f *is* $\hat{f}(\xi) = exp(-\xi^2/2)$ *so is of the form* $\exp(\phi(\xi))$ *with* $\phi(\xi) = -\xi^2/2$. *Then,*

$$\phi(\alpha\xi) + \phi(n\xi) = -\frac{\alpha^2\xi^2}{2} - \frac{n^2\xi^2}{2} = -\frac{(\alpha^2+n^2)\xi^2}{2}.$$

Thus, $h(\alpha, n) = \sqrt{\alpha^2 + n^2}$ *and*

$$\lim_{n\to\infty} \frac{\sqrt{\alpha^2 + n^2}}{n} = 1.$$

So, from Proposition 5.3.2, $\{\pi_n(\theta|x)\}_n$ *converges expected logarithmically to the formal prior* $\pi(\theta|x)$ *obtained for* $\pi(\theta) = 1$.

– *Consider the Cauchy model, then we have* $f(t) = \frac{1}{\pi(1+t^2)}$. *The corresponding sequence of priors* $\{\Pi_n\}_n$ *given by Proposition 5.3.2 is the sequence of Cauchy distributions with location parameter* 0 *and scale parameter* n. *The Fourier transform of* f *is* $\hat{f}(\xi) = exp(-|\xi|)$ *so is of the form* $\exp(\phi(\xi))$ *with* $\phi(\xi)) = -|\xi|$. *Then,*

$$\phi(\alpha\xi) + \phi(n\xi) = -|\alpha\xi| - |n\xi| = -|\xi|(|\alpha| + |n|).$$

Thus, $h(\alpha, n) = |\alpha| + |n|$ *and*

$$\lim_{n\to\infty} \frac{|\alpha| + |n|}{n} = 1.$$

So, from Proposition 5.3.2, $\{\pi_n(\theta|x)\}_n$ *converges expected logarithmically to the formal prior* $\pi(\theta|x)$.

Bibliography

Bailey, L. L., Simons, T. R., and Pollock, K. H. (2004). Comparing population size estimators for plethodontid salamanders. *Journal of Herpetology*, 38(3):370–380.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons Ltd., Chichester.

Bauer, H. (2001). *Measure and integration theory*, volume 26 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin. Translated from the German by Robert B. Burckel.

Bedrick, E. J. (1994). Maximum-Likelihood Estimation for the Removal Method. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 22(2):285–293.

Berger, J. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *J. American Statist. Assoc.*, 95:1269–1277.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. In *Bayesian statistics, 4 (Peñíscola, 1991)*, pages 35–60. Oxford Univ. Press, New York.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.*, 37(2):905–938.

Bernardo, J.-M. (1979a). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147.

Bernardo, J.-M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147. With discussion.

Bernardo, J.-M. (1997). Noninformative priors do not exist: A discussion. *Journal of Statistical Planning and Inference*, 65:159–189.

Bernardo, J.-M. and Smith, A. F. M. (1994). *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E., and Mekhnacha, K. (1998a). Interprétation ou description (i): Proposition pour une théorie probabiliste des systèmes cognitifs sensi-moteurs. *Intellectica*, 26-27:257–311.

Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E., and Mekhnacha, K. (1998b). Interprétation ou description (ii): Fondements mathématiques de l'approche f+d. *Intellectica*, 26-27:313–336.

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

Billingsley, P. (1986). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition.

Bioche, C. and Druilhet, P. (2015). Approximation of improper prior by vague priors. *To appear in Bernoulli*.

Blackwell, D. (1951). On the translation parameter problem for discrete variables. *Ann. Math. Statistics*, 22:393–399.

Bohrmann, T. F., Christman, M. C., and Smith, S. J. (2012). Evaluating sampling efficiency in depletion surveys using hierarchical Bayes. *Canadian Journal of Fisheries and Aquatic Sciences*, 69(6):1080–1090.

Bolfarine, H., Leite, J. G., and Rodrigues, J. (1992). On the Estimation of the Size of a Finite and Closed Population. *Biometrical Journal*, 34(5):577–593.

Bord, S., Druilhet, P., Gasqui, P., Abrial, D., and Vourc'h, G. (2014). Bayesian estimation of abundance based on removal sampling under weak assumption of closed population with catchability depending on environmental conditions. Application to tick abundance. *Ecological Modelling*, 274(0):72–79.

Bourbaki, N. (1971). *Éléments de mathématique. Topologie générale. Chapitres 1 à 4*. Hermann, Paris.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont. Addison-Wesley Series in Behavioral Science: Quantitative Methods.

Brun, M., Abraham, C., Jarry, M., Dumas, J., Lange, F., and Prevost, E. (2011). Estimating an homogeneous series of a population abundance indicator despite changes in data collection procedure: A hierarchical Bayesian modelling approach. *Ecological Modelling*, 222(5):1069–1079.

Carle, F. L. and Strub, M. R. (1978). A New Method for Estimating Population Size from Removal Data. *Biometrics*, 34(4):621–630.

Chan, Y., Anderson, C., and Hadly, E. (2006). Bayesian estimation of the timing and severity of a population bottleneck from ancient dna. *PLoS Genetics*, 2.

Chatterjee, N. D., Krüger, R., Haller, G., and Olbricht, W. (1998). The Bayesian approach to an internally consistent thermodynamic database: theory, database, and generation of phase diagrams. *Computer.*, 133:149–168.

Chernoff, H. (1956). Large-sample theory: parametric case. *Ann. Math. Statist.*, 27:1–22.

Cousins, R. D. (1995). Why isn't every physicist a Bayesian? *Amer. J. Phys.*, 63(5):398–410.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158.

Dauphin, G., Prevost, E., Adams, C. E., and Boylan, P. (2009). A Bayesian approach to estimating Atlantic salmon fry densities using a rapid sampling technique. *Fisheries Management and Ecology*, 16(5):399–408.

Dauxois, J.-Y., Druilhet, P., and Pommeret, D. (2006). A Bayesian choice between Poisson, binomial and negative binomial models. *Test*, 15(2):423–432.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B*, 35:189–233.

Demortier, L. (2006). Bayesian reference analysis. In Lyons, L. and Ünel, M. K., editors, *Statistical problems in particle physics, astrophysics and cosmology*, pages 11–+. Imp. Coll. Press, London.

Deneve, S. (2005). Bayesian inference in spiking neurons. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 353–360. MIT Press, Cambridge, MA.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.*, 7(2):269–281.

Dodd, C. K. and Dorazio, R. M. (2004). Using counts to simultaneously estimate abundance and detection probabilities in a salamander community. *Herpetologica*, 60(4):468–478.

Dorazio, R. M. and Jelks, H. L. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics*, 61(4):1093–1101.

Dorazio, R. M., Royle, J. A., Soderstrom, B., and Glimskar, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854.

Druilhet, P. and Pommeret, D. (2012). Invariant conjugate analysis for exponential families. *Bayesian Anal.*, 7(4):903–916.

Eaton, M. L. (1989). *Group invariance applications in statistics*. NSF-CBMS Regional Conference Series in Probability and Statistics, 1. Institute of Mathematical Statistics, Hayward, CA.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, 7(6):509–520.

Fisher, R. (1922). On the mathematical foundations of theoretical Statistics. *Philos. Trans. Roy. Soc. London*, 222:309–368.

Fisher, R. (1930). Inverse probability. *Proc. Cambridge Philos. Soc.*, 26:528–535.

Fisher, R. (1935). The fiducial argument in statistical inference. *Annals. of Eugenics*, 6:391–8.

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Fraser, D. A. S., Monette, G., and Ng, K. W. (1985). Marginalization, likelihood and structured models. In *Multivariate analysis VI (Pittsburgh, Pa., 1983)*, pages 209–217. North-Holland, Amsterdam.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics*. Constable, London. Reprinted by Dover, New York, 1960.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin & Co., Ltd., London; Hafner Publishing Co., New York, N. Y.

Good, I. J. (1969). What is the use of a distribution? In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 183–203. Academic Press, New York.

Gove, J. H., Linder, E., and Tzilkowski, W. M. (1995). Biomodality of the combined removal and signs-of-activities estimator for sampling closed animal populations. *Environmental and Ecological Statistics*, 3(1):65–78.

Greenleaf, F. P. (1969). *Invariant means on topological groups and their applications*. Van Nostrand Mathematical Studies, No. 16. Van Nostrand Reinhold Co., New York-Toronto, Ont.-London.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(01):55–61.

Hartigan, J. (1964). Invariant prior distributions. *Ann. Math. Statist.*, 35:836–845.

Hartigan, J. A. (1983). *Bayes theory*. Springer Series in Statistics. Springer-Verlag, New York.

Hartigan, J. A. (1996). Locally uniform prior distributions. *Ann. Statist.*, 24(1):160–173.

Hayne, D. W. (1949). An Examination of the Strip Census Method for Estimating Animal Populations. *The Journal of Wildlife Management*, 13(2):pp. 145–157.

Heath, D. and Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.*, 17(2):907–919.

Heyer, W., Donnely, M., McDiarmid, R., Hayek, L.-A. C., and Foster, M. S. (1994). *Measuring and monitoring biological diversity: standard methods for amphibians.* Smithsonian Institution Press.

Hirst, D. (1994). An Improved Removal Method for Estimating Animal Abundance. *Biometrics*, 50(2):501–505.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev. (2)*, 106:620–630.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics*, 4:227–291.

Jaynes, E. T. (1980). Marginalization and prior probabilities. In Zelner, A., editor, *Bayesian Analysis in Econometrics and Statistics*. North-Holland, Amsterdam.

Jaynes, E. T. (2003). *Probability theory*. Cambridge University Press, Cambridge. The logic of science, Edited and with a foreword by G. Larry Bretthorst.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London. Ser. A.*, 186:453–461.

Jeffreys, H. (1961). *Theory of probability*. Third edition. Clarendon Press, Oxford.

Kahn, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *Amer. Statist.*, 41(1):38–40.

Kakutani, S. (1948). On equivalence of infinite product measures. *Ann. of Math. (2)*, 49:214–224.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of American Statistical Association*, 91(435):1343–1370.

Kording, K. P. (2004). Bayesian integration in sensorimotor learning. *Nature*, 15 (427):244–7.

Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, 22:79–86.

Lane, D. A. and Sudderth, W. D. (1983). Coherent and continuous inference. *Ann. Statist.*, 11(1):114–120.

Lang, S. (1977). *Analyse réelle*. InterEditions, Paris.

Laplace, P. S. (1786). Sur les Naissances, les Mariages et les Morts Histoire de L'Academic Royale des Sciences.

Laplace, P.-S. (1995). *Théorie analytique des probabilités. Vol. II*. Éditions Jacques Gabay, Paris. Reprint of the 1820 third edition (Book II) and of the 1816, 1818, 1820 and 1825 originals (Supplements).

Lebeltel, O., Bessière, P., Diard, J., and Mazer, E. (2003). Bayesian robots programming. *Autonomous Robots*, 16(1):49–79.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Leslie, P. H. and Davis, D. H. S. (1939). An Attempt to Determine the Absolute Number of Rats on a Given Area. *Journal of Animal Ecology*, 8(1):94–113.

Lindley, D. V. (1990). The 1988 Wald Memorial Lectures: the present position in Bayesian statistics. *Statist. Sci.*, 5(1):44–89. With comments and a rejoinder by the author.

MacKenzie, D. and Royle, J. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, 44(6):1105–1114.

Mantyniemi, S., Romakkaniemi, A., and Arjas, E. (2005). Bayesian removal estimation of a population size under unequal catchability. *Canadian Journal of Fisheries and Aquatic Sciences*, 62(2):291–300.

Moran, P. A. P. (1951). A Mathematical Theory of Animal Trapping. *Biometrika*, 38(3/4):307–311.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). *Statistical inference from capture data on closed animal populations*, volume 62. wildlife society.

Pohorille, A. and Darve, E. (2006). A Bayesian approach to calculating free energies in chemical and biological systems. In *Bayesian inference and maximum entropy method in science and engineering*, volume 872, pages 23–30.

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107:1–97.

Pouget, A., Dayan, P., and Zemel, R. S. (2003). Inference and computation with population codes. *Annu Rev Neurosci*, 26:381–410.

Rényi, A. (1970). *Foundations of probability*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam.

Rivot, E., Prevost, E., Cuzol, A., Bagliniere, J.-L., and Parent, E. (2008). Hierarchical Bayesian modelling with habitat and time covariates for estimating riverine fish population size by successive removal method. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(1):117–133.

Robert, C. P. (2007). *The Bayesian choice*. Springer Texts in Statistics. Springer, New York, second edition. From decision-theoretic foundations to computational implementation.

Royle, J. A. (2004a). Modeling Abundance Index Data from Anuran Calling Surveys Modelaje de Datos de Índices de Abundancia a partir de Muestreos de Llamados de Anuros. *Conservation Biology*, 18(5):1378–1385.

Royle, J. A. (2004b). N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics*, 60(1):108–115.

Royle, J. A. and Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):249–263.

Schnute, J. (1983). A new approach to estimating populations by the removal sampling method. *Canadian Journal of Fisheries and Aquatic Sciences*, 40(12):2153–2169.

Schwarz, C. J. and Seber, G. A. F. (1999). Estimating animal abundance: Review III. *Statistical Science*, 14(4):427–456.

Seber, G. A. F. (1982). The estimation of animal abundance: Griffin.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3:Art. 3, 29 pp. (electronic).

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 197–206. University of California Press, Berkeley and Los Angeles.

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. *Bernoulli, Bayes, Laplace*, Anniversary Volume:217–240.

Stone, M. (1963). The posterior t distribution. *Ann. Math. Statist*, 34:568–573.

Stone, M. (1964). Comments on a posterior distribution of Geisser and Cornfield. *J. Roy. Statist. Soc. Ser. B*, 26:274–276.

Stone, M. (1965). Right Haar measure for convergence in probability to quasi posterior distributions. *Ann. Math. Statist.*, 36:440–453.

Stone, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.*, 41:1349–1353.

Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.*, 71(353):114–125. With comments by James M. Dickey, John W. Pratt, D. V. Lindley, George A. Barnard, G. E. P. Box and G. C. Tiao, D. A. S. Fraser and C. Villegas and a reply by the author.

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, 59:369–375.

Taraldsen, G. and Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *Ann. Statist.*, 41(1):323–341.

Taraldsen, G. and Lindqvist, B. H. (2015a). Conditional probability and improper priors. *Commun. Stat.A-Theor (accepted)*.

Taraldsen, G. and Lindqvist, B. H. (2015b). Fiducial and posterior sampling. *Commun. Stat.A-Theor.*

Taraldsen, G. and Lindqvist, H. (2010). Improper priors are not improper. *The American Statistician*, 64(2):154–158.

Tuyl, F., Gerlach, R., and Mengersen, K. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Anal.*, 4(1):151–158.

Villegas, C. (1967). On qualitative probability. *Amer. Math. Monthly*, 74:661–669.

Vines, K. S., Evilia, R. F., and Whittenburg, S. L. (1993). Bayesian analysis investigation of chemical exchange above and below the coalescence point. *Journal of physical chemistry*, 97:4941–4944.

Wallace, D. L. (1959). Conditional confidence level properties. *Ann. Math. Statist.*, 30:864–876.

White, G. C., Leffler, B., and Laboratory, L. A. N. (1982). *Capture-recapture and removal methods for sampling closed populations*. LA-8787-NERP. Los Alamos National Laboratory.

Wilkinson, D. J. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Brief bioinform.*

Wilkinson, G. (1971). In discussion of Godambe, V. P. and Thompson, Mary e. (1971). Bayes, fiducial and frequency aspects of statistical inference in regression analysis in survey-sampling. *J. Roy. Statist. Soc. Ser. B*, 33:361–390.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002a). *Analysis and Management of Animal Populations*. Academic Press, San Diego, USA & London, UK, 1st edition.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002b). *Analysis and management of animal populations : modeling, estimation and decision making*. Academic Press, San Diego, Calif.

Wu, S., Chen, D., Niranjan, M., and Amari, S. I. (2003). Sequential Bayesian decoding with a population of neurons. *Neural Comput*, 15(5):993–1012.

Wyatt, R. J. (2002). Estimating riverine fish population size from single- and multiple-pass removal sampling using a hierarchical model. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(4):695–706.

Zippin, C. (1956). An Evaluation of the Removal Method of Estimating Animal Populations. *Biometrics*, 12(2):163–189.

Zippin, C. (1958). The Removal Method of Population Estimation. *The Journal of Wildlife Management*, 22(1):pp. 82–90.

Approximation d'*a priori* impropres et applications

Résumé : Le but de cette thèse est d'étudier l'approximation d'*a priori* impropres par des suites d'*a priori* propres. Nous définissons un mode de convergence sur les mesures de Radon strictement positives pour lequel une suite de mesures de probabilité peut admettre une mesure impropre pour limite. Ce mode de convergence, que nous appelons *convergence q-vague*, est indépendant du modèle statistique. Il permet de comprendre l'origine du paradoxe de Jeffreys-Lindley. Ensuite, nous nous intéressons à l'estimation de la taille d'une population. Nous considérons le modèle du removal sampling. Nous établissons des conditions nécessaires et suffisantes sur un certain type d'*a priori* pour obtenir des estimateurs *a posteriori* bien définis. Enfin, nous montrons à l'aide de la convergence q-vague, que l'utilisation d'*a priori* vagues n'est pas adaptée car les estimateurs obtenus montrent une grande dépendance aux hyperparamètres.

Mots-clés : *A priori* conjugués, *a priori* de référence, *a priori* impropres, *a priori* non-informatifs, *a priori* vagues, convergence d'*a priori*, convergence logarithmique, paradoxe de Jeffreys-Lindley, removal sampling, statistiques bayésiennes.

Approximation of improper priors and applications

Abstract: The purpose of this thesis is to study the approximation of improper priors by proper priors. We define a convergence mode on the positive Radon measures for which a sequence of probability measures could converge to an improper limiting measure. This convergence mode, called q-vague convergence, is independant from the statistical model. It explains the origin of the Jeffreys-Lindley paradox. Then, we focus on the estimation of the size of a population. We consider the removal sampling model. We give necessary and sufficient conditions on the hyperparameters in order to have proper posterior distributions and well define estimate of abundance. In the light of the q-vague convergence, we show that the use of vague priors is not appropriate in removal sampling since the estimates obtained depend crucially on hyperparameters.

Key-words: Bayesian statistic, conjugate prior, convergence of priors, improper prior, Jeffreys-Lindley paradox, logarithmic convergence, noninformative prior, reference prior, removal sampling, vague prior.